THE ROYAL SOCIETY

Exact solutions to the transportation problem on the line

Dedicated to Elliott H. Lieb in honour of his 65th birthday; his exact results in one dimension continue to inspire.

By Robert J. McCann†

Department of Mathematics, Brown University, Providence, RI 02912, USA and Institut des Hautes Etudes Scientifiques, 91440 Bures-sur-Yvette, France

Received 5 January 1998; accepted 6 April 1998

Given distributions μ of production and ν of consumption on the line, the Monge–Kantorovich problem is to decide which producer should supply each consumer in order to minimize the total transportation costs. Here cost will be assumed to be a strictly concave function of the distance, which translates into an economy of scale for longer trips and may encourage cross-hauling. The resulting solutions display a hierarchical structure that reflects a striking separation into local and global scales also found in the real world. Moreover, this structure can be exploited to reduce the infinite-dimensional linear problem to a convex minimization in m variables, where $2m+2$ counts the number of times that $\mu-\nu$ changes sign. A combinatorial algorithm is then derived which yields exact solutions by optimizing a certain finite sequence of convex, separable network flows.

Keywords: Monge–Kantorovich; spatial economics; trnsportation; optimal map; network flow optimization; convex programming; hierarchical structure

1. Introduction

In the classical transportation problem, one is given a distribution μ of iron mines throughout the countryside, and a distribution ν of factories that require iron ore, and asked to decide which mines should supply ore to each factory in order to minimize the total transportation costs. The cost per ton of ore transported from the mine at x to the factory at y is given by a function $c(x, y)$, so the problem can be formulated as a linear program. Indeed, the question helped to motivate the development of duality theory by Kantorovich (1942) and Koopmans (1949), though its origins date much further back to Monge (1781). The present paper concerns itself with the solution to this problem in the special case of mines and factories that are distributed continuously along the line, with a cost $c(x, y) = h(|x-y|)$ given by a strictly concave function $h \geqslant 0$ of the distance.

Although somewhat idealized, the setting just described provides a reasonable model for applications in which shipping occurs along a single route: a railway line or highway, or along one coast of North America. Concavity of h reflects a shipping tariff that increases with the distance, even while the cost per mile shipped goes down.

† Present address: Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 3G3.

Despite its economic relevance, transportation with concave costs has received much less attention than the same problem for convex costs. The latter enjoys a sizable literature and, at least in one dimension, has been completely understood (see Tchen (1980) or Rachev (1984, § 2.3) for reviews). For concave costs on the other hand, it was only recently observed by Gangbo & McCann (1996) that the solutions

> will not be smooth, but display an intricate structure which—for us—was unexpected; it seems equally fascinating from the mathematical and the economic point of view.... To describe one effect in economic terms: the concavity of the cost function favours a long trip and a short trip over two trips of average length; as a result, it can be efficient for two trucks carrying the same commodity to pass each other travelling opposite directions on the highway: one truck must be a local supplier, the other on a longer haul. In optimal solutions, such 'pathologies' may nest on many scales, leading to a natural hierarchy among the regions of supply [where $\mu \geqslant \nu$] and of demand [where $\mu \leqslant \nu$].

Our purpose here is to expose the nature of this hierarchy, exploring its theoretical and computational implications. Its very existence suggests a simple explanation for the emergence of spatial price and distribution patterns on diverse scales. Specific features of these patterns are predicted that are not sensitive to the details of the cost, and could therefore be tested against observations. Finally, the local/global structure of the hierarchy is exploited to reduce the infinite-dimensional optimization problem—with its continuously distributed production excess—to a minimization in finitely many (say m) variables; here $2m+2$ counts the number of times that the density of $\mu - \nu$ changes sign along the line. This finite-dimensional problem can in turn be solved by a combinatorial sequence of optimizations of convex, separable network flows.

To formulate the problem mathematically, take the distributions μ and ν of production and consumption along the line to belong to $\mathcal{M}_+(\mathbb{R})$, the space of non-negative (Borel) measures with finite total mass. Equality of net supply with net demand is enforced by taking $\mu[\mathbb{R}] = \nu[\mathbb{R}]$, and the problem is then to minimize the *transport cost*,

$$\mathcal{C}(\gamma) := \int_{\mathbb{R}^2} c(x,y)\,\mathrm{d}\gamma(x,y), \tag{1.1}$$

among non-negative measures γ on the plane that have μ and ν for *marginals*: $\mu[U] = \gamma[U \times \mathbb{R}]$ and $\gamma[\mathbb{R} \times U] = \nu[U]$ for every (Borel) set $U \subset \mathbb{R}$. The collection of such γ is a convex subset of $\mathcal{M}_+(\mathbb{R}^2)$, which will be denoted here by $\Gamma(\mu,\nu)$.

It is well known that the linear functional $\mathcal{C}(\gamma)$ attains its minimum on $\Gamma(\mu,\nu)$ as long as $c : \mathbb{R}^2 \longrightarrow [0,\infty]$ is lower semi-continuous (see, for example, Kellerer 1984, thm 2.19). We therefore say a measure γ is *optimal* when its transport cost is a minimum among all measures in $\mathcal{M}_+(\mathbb{R}^2)$ with the same marginals as γ. To avoid trivialities, we also insist $\mathcal{C}(\gamma) < \infty$ if γ is called optimal. For us the object of geometrical interest will be the *support* $\operatorname{spt}\gamma$ of the optimal measure, meaning the smallest closed subset of $\mathbb{R}^2$ carrying full mass for γ. For example, strictly *convex* costs $c(x,y) = \ell(x-y)$ imply $\operatorname{spt}\gamma$ is non-decreasing in the plane: i.e. the left-most mines supply the left-most factories (Appendix A). More generally, when γ is optimal, $(x,y) \in \operatorname{spt}\gamma$ means it is efficient to transport from the mine at x to the factory at y.

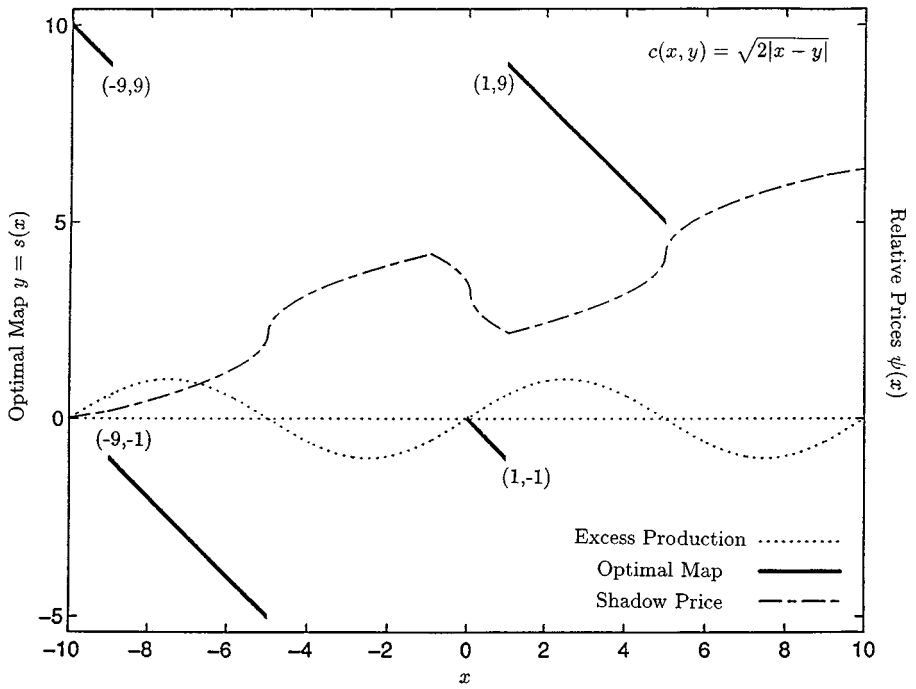

Figure 1. Optimal map and shadow prices for a sinusoidal distribution of mass.

For heuristic purposes, it is useful to understand two of the main results of Gangbo & McCann (1996): as long as μ concentrates no mass on any single point $x \in \mathbb{R}$, nor on $\operatorname{spt}\nu$, then for costs $c(x,y) = h(|x-y|)$ given by strictly concave functions $h \geqslant 0$ it remains true that (i) the optimal measure γ is unique in $\Gamma(\mu,\nu)$; and (ii) there is a mapping $s : \mathbb{R} \longrightarrow \mathbb{R}$ whose graph, $\{(x,s(x))\}$, carries full mass for γ. In other words, it is optimal for almost every mine x to ship its output to a single factory: $\operatorname{spt}\gamma$ can be thought of as a map from μ to ν. This map s is uniquely determined μ-almost everywhere, and contains enough information to reconstruct the optimal measure γ from the formula $\gamma[\Omega] = \mu[\{x \mid (x,s(x)) \in \Omega\}]$ which holds for $\Omega \subset \mathbb{R}^2$. It will be called the *optimal map* between μ and ν. An example serves to illustrate.

Example 1.1 (Local and global supply). Distribute excess production $\rho = \mu - \nu$ sinusoidally over -10 to 10: $\mathrm{d}\rho(x) = \sin(\frac{1}{5}\pi x)\,\mathrm{d}x$. For the cost

$$c(x,y) := \sqrt{2|x-y|},$$

the optimal map between $\mu = \rho_+$ and $\nu = \rho_-$ is given by figure 1:

$$s(x) := \begin{cases} -x - 10, & \text{where } -9 < x < -1, \\ -x, & \text{where } |x| < 1 \text{ or } |x| > 9, \\ -x + 10, & \text{where } 1 < x < 9. \end{cases} \tag{1.2}$$

The same map is represented schematically in figure 2 by semi-circular arcs connecting each $x \in \operatorname{spt}\mu$ to its destination $s(x) \in \operatorname{spt}\nu$. Our key observation about the structure of this map, which holds for any pair of measures and strictly concave cost functions on the line, is that two arcs never cross (compare the patterns

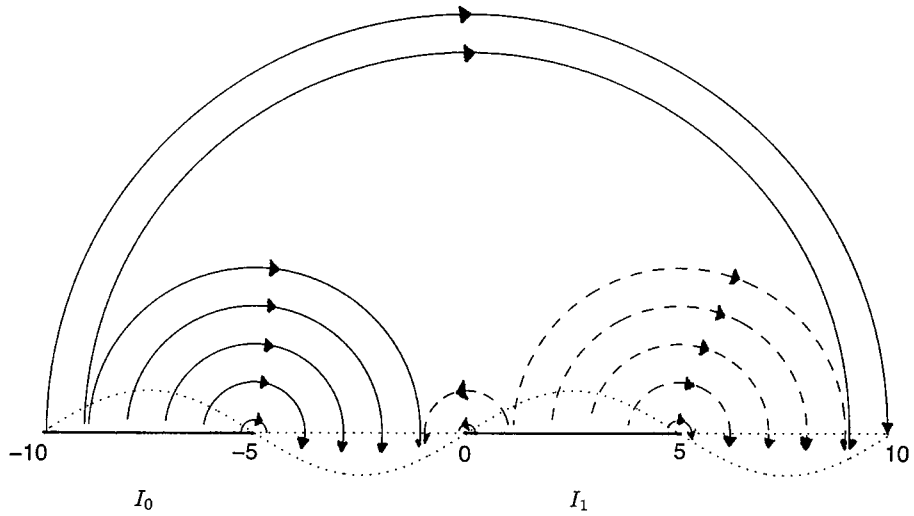

Figure 2. The no-crossing rule: I_1 supplies locally relative to I_0.

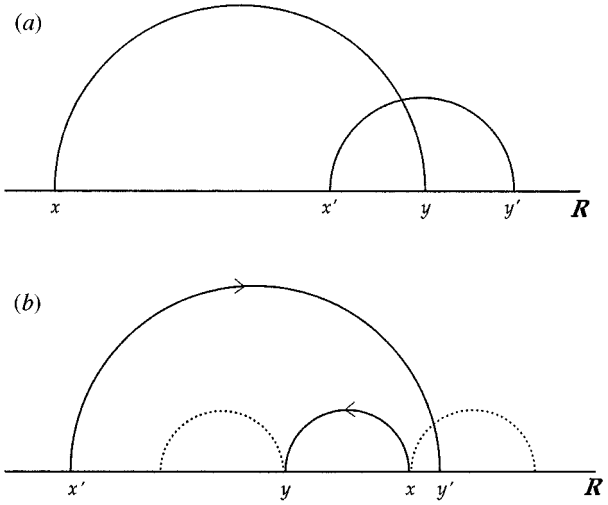

Figure 3. Forbidden patterns: mine x will not supply factory y if mine x' supplies factory y'; (a) violates the no-crossing rule; (b) violates the rule of three.

in figure 2 with those excluded by figure 3a). The optimal map is piecewise non-increasing as one consequence, while the hierarchy illustrated by figure 2, where $I_1 := [0, 5]$ seems to supply locally with respect to $I_0 := [-10, -5]$, is another. From this no-crossing rule, one also infers that each arc joining x to $s(x)$ encircles zero net mass: $\rho[[x, s(x)]] = 0$. The only decisions left to make are where the jump increases in s occur; in our example these discontinuities are selected by the equation

$$c(1, -1) + c(-9, 9) = c(1, 9) + c(-9, -1). \tag{1.3}$$

More generally, when a finite union, $\bigcup_{i=0}^{m} I_i$, of intervals contains the full mass of μ but no mass of ν, then the maps that do not have crossings can be parametrized by m variables: essentially the fractions $\phi_i \in [0, 1]$ of mass to be transported to

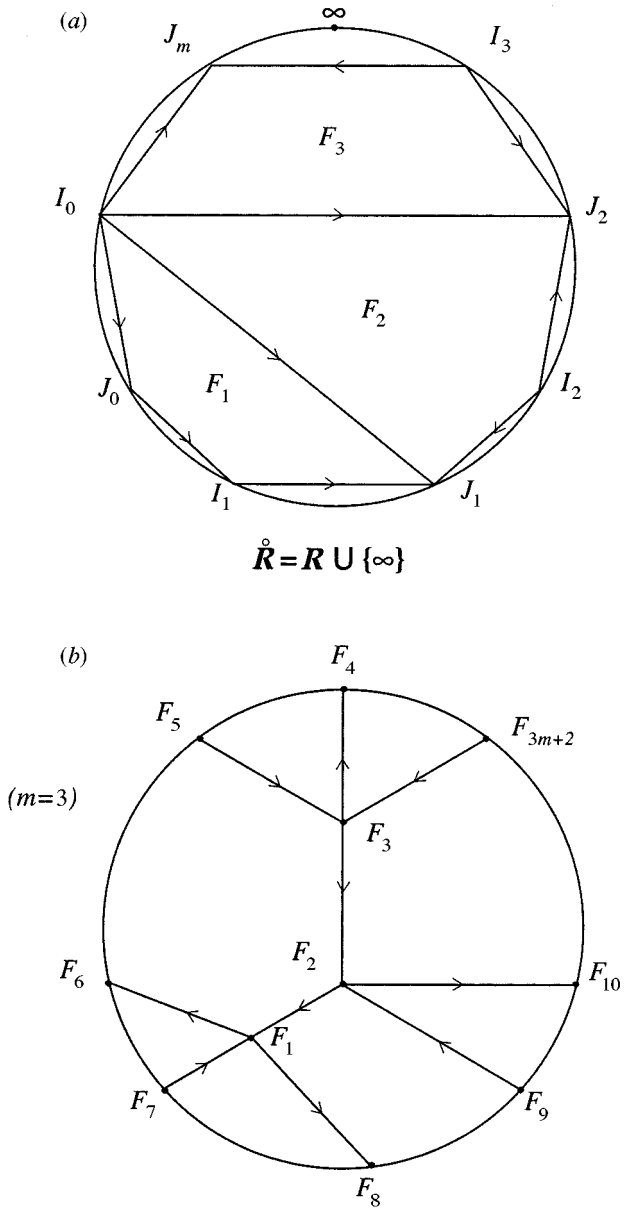

Figure 4. (*a*) An uncrossed network G. (b) The dual network G^*.

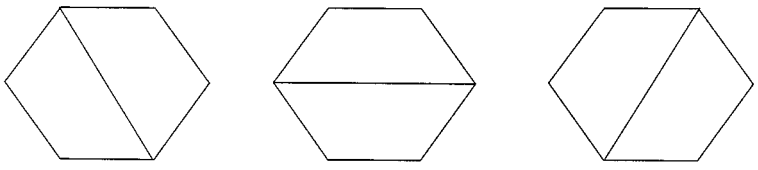

Figure 5. All three uncrossed networks on six nodes ($m = 2$).

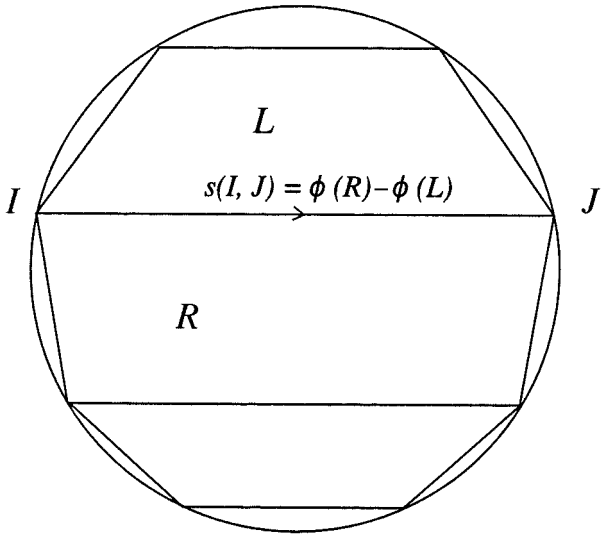

Figure 6. The 12 uncrossed network on eight nodes ($m = 3$) are given by this diagram in four orientations together with eight orientations of figure 4*a*.

the left from each interval $I_i \subset \mathbb{R}$ of supply. This parametrization is introduced in § 3, where it is also pointed out that certain combinatorial data beyond these mass fractions are required to reconstruct a unique map: one needs to know which intervals I_i supply 'locally' with respect to the others. These combinatorial data are visualized by identifying the $m + 1$ intervals, numbered from left to right along the line, with the odd vertices of a regular $(2m + 2)$-sided polygon G. The even vertices represent the intervals separating I_i from I_{i+1}; these complementary intervals are denoted here by J_i. The combinatorial freedom is then accounted for by introducing enough additional edges between existing vertices to subdivide the polygon's interior into quadrilaterals, as in figure 4*a*. Since no two edges may cross each other, this division can be achieved in $d_m = (3m \text{ choose } m)/(2m + 1)$ ways†. After such a division is specified, one searches for the optimal mapping $s : \bigcup I_i \longrightarrow \mathbb{R} \setminus \bigcup I_i$, which is free from crossings and compatible with the *network* G, in the sense that no mass is transported from I_i to J_j unless G includes an arc connecting those vertices.

In § 4 we show such maps to be parametrized by a convex polytope $\Phi_G \subset \mathbb{R}^m$. Surprisingly, in the mass coordinates $(\phi_1, \ldots, \phi_m) \in \Phi_G$, the transport cost (1.1) turns out to be a convex function of the form

$$\mathcal{C}_G(\phi_1, \ldots, \phi_m) = \sum_{i=1}^{m} C_i(\phi_i). \tag{1.4}$$

Thus, the minimization of $\mathcal{C}(\gamma)$ on $\Gamma(\mu, \nu)$ is reduced to solving d_m optimal-flow problems on a special set of networks G. Each variable ϕ_i controls the location of the spatial boundary between two regions in the local/global hierarchy. In example 1.1, where $m = 1$, the equilibrium equation $\mathrm{d}\mathcal{C}_G/\mathrm{d}\phi_1 = 0$ turns out to be equivalent to (1.3). More generally, *separability* (1.4) of the cost $\mathcal{C}_G$ affords efficient computations (see Rockafellar 1984).

† For $m = 2$ and $m = 3$ the possibilities are enumerated in figures 5 and 6, while the formula for d_m is due to Erdélyi & Etherington (1940).

A final section probes the relationship between the different networks G. For optimality, it proves to be enough that a map or measure minimizes the transport cost only on those networks with which it is compatible. Equilibrium with respect to perturbations of $(\phi_1, \ldots, \phi_m)$ is therefore sufficient to assure global stability. Verifying optimality then becomes a finite calculation, while finding a global minimum need not require consideration of all $(3m \text{ choose } m)/(2m+1)$ possible networks G. This theorem hints at the presence of some underlying convexity, not only within each network problem, but among the different networks G. Its proof hinges on a remarkable observation, stated here as proposition 5.5. This proposition provides a consistency condition for gluing together optimal maps $s_1 : I \longrightarrow I$ and $s_2 : \tilde{I} \longrightarrow \tilde{I}$ defined on complementary intervals to obtain an optimal map throughout the line: namely, that it suffices for s_1 to agree with s_2 at the boundary points separating I from $\tilde{I} := \mathbb{R} \setminus I$. As long as this matching condition is satisfied, the interval I is decoupled from the rest of the line. At least in one dimension, this confirms the intuition that efficient distribution may be achieved by optimizing independently on different scales, e.g. local, regional, national, before allowing competition to adjust the boundaries between scales. Whether such a hierarchy persists in more than one dimension remains purely speculative; perhaps the trace of such a structure is suggested by Gangbo & McCann (1996, fig. 2*d*).

(*a*) *References to related works*

The transportation problem was first studied by Monge (1781): his formulation is in terms of volume-preserving maps $s : U \longrightarrow V$ between two subsets $U, V \subset \mathbb{R}^3$ of equal volume, and he measured optimality of these maps against Euclidean distance $c(x, y) = |x - y|$. The formulation we have used, in terms of joint measures with given marginals, is due to Kantorovich (1942), who also discovered the existence of a dual problem when μ and ν measure an abstract space metrized by $c(x, y)$; this dual problem involves potentials that play the role of the shadow prices introduced by Koopmans (1949). One year earlier, Hitchcock (1941) had outlined an algorithm, akin to the simplex method of Dantzig (1951), for optimizing transportation between finitely many mines and factories. A variation of the problem in which revenue, μ, must be determined for two competitors $x, x' \in \mathbb{R}$ offering goods to a market, ν, spread out along the line had been studied by Hotelling (1929), who offered many interesting interpretations for location.

A continuous-flow model for transportation in two dimensions was proposed by Beckmann (1952), and subsequently developed by Beckmann & Puu (1985). While this model offers much flexibility, the solution concept differs markedly from the present one: instead of a mapping from the plane to itself, a solution is given by a vector field representing continuous flow of ore from mines to factories. This flow has a definite direction and velocity at each point, so the cross-hauling illustrated by the arrows in figure 2 is precluded from their formulation: the Beckmann & Puu (1985) model predicts only the average flow through each point on the plane without distinguishing individual shipment origins or destinations. An advantage of their model is that transport costs may be nonlinear in quantity shipped, and distances non-Euclidean, though they are limited to costs depending linearly on the distance. The central result of their theory, as of the Kantorovich and Koopmans work, is that optimal flow is determined by the existence of a consistent shadow pricing scheme.

Although duality is not the focus of the present study, the shadow price $\psi(x)$ along the line may be determined from the optimal map through the equations

$$\psi'(x) = -\frac{\partial c}{\partial x}(x, s(x)) \ [\mu \text{ a.e.}] \quad \text{and} \quad \psi'(y) = \frac{\partial c}{\partial y}(s^{-1}(y), y) \ [\nu \text{ a.e.}]. \tag{1.5}$$

Thus, up to an overall constant, the equilibrium prices in example 1.1 are given by translates and shifts of a fixed function reproduced on various intervals (figure 1)

$$\psi(x) := \begin{cases} |x+5|^{-1/2}(x+5) - 1, & \text{where } -9 < x < -1, \\ \pm|x|^{-1/2}x, & \text{where } |x| > 9 \text{ or } |x| < 1, \\ |x-5|^{-1/2}(x-5) + 1, & \text{where } 1 < x < 9, \end{cases}$$

the price difference between the head and tail of every arc in figure 2 agreeing with the transport cost: $\psi(s(x)) = \psi(x) + c(x, s(x))$.

A history of the Monge–Kantorovich problem and its applications to statistics is outlined in Rachev (1984), where many references may also be found. Sources for more recent developments include the articles of Rüschendorf (1991), Evans & Gangbo (1999) and Gangbo & McCann (1996). The former gives a characterization of optimality that applies to abstract costs and measure spaces, while the latter establish existence and uniqueness of optimal maps for convex or concave cost functions of Euclidean distance on $\mathbb{R}^n$.

Particularly germane to our concerns will be a theorem due to Smith & Knott (1992) that characterizes optimal measures via *c-cyclical monotonicity* of their support (see theorem 5.3). Originally derived by them from the duality based work of Rüschendorf (1991), this characterization appears with hindsight as a natural extension of the necessary condition (2.1) observed by Monge. It is interesting to note that c-cyclical monotonicity was discovered earlier (and independently) in a different economic context: it was introduced by Rochet (1987) to characterize incentive contracts offered to agents by a principal. Indeed, the mathematical structure of the principal-agent problem is closely intertwined with that of mass transport (a connection currently being pursued in joint work with Ivar Ekeland).

On the line, the transportation problem has been studied by many authors: the contributions of Gini, Hoeffding, Salvemini, Fréchet, Dall'Aglio, Cambanis, Simons, and Stout are outlined in Tchen (1980) and Rachev (1984, § 2.3). For costs $c(x, y) = |x-y|^p$ with $p \geqslant 1$, the results summarized in Appendix A were stated by Dall'Aglio (1956), though special cases were anticipated by Hoeffding, Salvemini and Fréchet. The corresponding rearrangement inequalities go back at least as far (see Lorentz 1953), though conditions for strict inequality (and hence uniqueness) came much later (Lieb 1977). In the meanwhile, Bertino (1966) studied costs given by strictly concave functions of the distance, and pointed out that the non-decreasing map could not be optimal. Uniqueness of solution, though conjectured in the abstract, failed to be addressed in the text; under suitable hypotheses it was affirmed in Gangbo & McCann (1996). Finally, we have learned that in work concurrent with but independent of the present paper, Uckelmann (1997) has obtained exact solutions exhibiting singularities similar to those of figure 1. His method is an outgrowth of the work of Rüschendorf; it requires less structure from the cost but is restricted to measures μ and ν distributed uniformly throughout two given intervals. At the same time, Bagdasarov (1998) explored the maximization problem dual to (1.1) in connection with the Kolmogorov–Landau inequalities for functions $f(x)$ on the line.

He exposed many properties of the optimal price $\psi(x)$ that have analogues in our theory, using them to characterize sharp bounds on intermediate derivatives of f. His concave modulus of continuity ω takes the place of our cost function c.

It may be interesting to close by noting that the structure of uncrossed sets and measures introduced in §§ 2 and 3 share similarities with the 'earthquakes' used by Thurston (1986) to represent hyperbolic structures on the plane.

2. The no-crossing rule for concave costs

The primary goal of this section is to establish the no-crossing rule implied by concave costs for optimal maps, together with a constraint on nesting transportation in opposite directions, which we refer to as the *rule of three*. We then introduce more terminology, and proceed by exploring implications of these transport rules. In particular, the no-crossing rule is shown to determine the optimal measure $\gamma \in \Gamma(\mu, \nu)$ uniquely when the density of $\mu - \nu$ changes sign at most twice along the line. Apart from technicalities, γ is selected by insisting that the optimal map s, and its inverse s^{-1}, be simultaneously approximated by an orientation reversing homeomorphism of the circle $\boldsymbol{S}^1 \approx \mathring{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$.

Both the no-crossing rule and 'rule of three' hinge on a fact originally observed by Monge (1781): for (x, y) and (x', y') from $\operatorname{spt}\gamma$, optimality of γ implies

$$c(x, y) + c(x', y') \leqslant c(x, y') + c(x', y); \tag{2.1}$$

otherwise, it would be more efficient to pair x with y' and x' with y. For $c(x, y) = \ell(x - y)$ with ℓ strictly convex, (2.1) is exactly the inequality that implies $\operatorname{spt}\gamma$ non-decreasing (as in Appendix A). Note as well the relationship between Monge's inequality and the equilibrium condition (1.3) of example 1.1.

Here, we are interested in the implications of (2.1) for costs $c(x, y) := h(|x - y|)$ given by strictly concave functions $h \geqslant 0$ of the distance. Therefore, associate to $x, y \in \mathbb{R}$ the smallest circle $O(x, y)$ in the plane that crosses the horizontal axis at both $(x, 0)$ and $(y, 0)$. The *no-crossing rule* implied by (2.1) asserts that $O(x, y)$ does not cross $O(x', y')$, precluding the pattern shown in figure 3*a*. Furthermore, suppose the circle $O(x, y)$ is enclosed by $O(x', y')$, but the signs of $y - x$ and $y' - x'$ are not the same. We have described this situation—which occurs in figure 2 (the arrows above the origin indicate direction of transport) and figure 3*b*—as two trucks passing in opposite directions. The second part of the lemma states the *rule of three*: namely that a concentric circle three times as large as $O(x, y)$ is also enclosed by $O(x', y')$. This *quantitative* conclusion is independent of the choice of concave cost; it could be tested against empirical data to see how the predictions of our simple model compare with transportation problems solved by the market in the real world.

Lemma 2.1. *Let $c(x, y) := h(|x - y|)$ with $h : [0, \infty) \longrightarrow \mathbb{R} \cup \{-\infty\}$ strictly concave and increasing. If $x, y, x', y' \in \mathbb{R}$ satisfy $c(x, y) + c(x', y') \leqslant c(x, y') + c(x', y)$, then:*

(i) *the circles $O(x, y)$ and $O(x', y') \subset \mathbb{R}^2$ do not intersect unless $x = x'$ or $y = y'$;*

(ii) *if the circle $O(x', y')$ encloses $O(x, y)$ but $(y - x)(y' - x') < 0$, then it encloses (without touching) a concentric circle $O(2x - y, 2y - x)$ three times as large.*

Proof. (i) Interchanging x with y or $(x,y) \leftrightarrow (x',y')$ if necessary, one assumes x greater than or equal to the remaining three numbers y, x', y' without losing generality. To produce a contradiction, assume the circles $O(x,y)$ and $O(x',y')$ intersect though $x \neq x'$ and $y \neq y'$. The only orderings consistent with these assumptions are (a) $x > x' \geqslant y > y'$; and (b) $x \geqslant y' > y \geqslant x'$. The second possibility (b) is easy to exclude: strict monotonicity of h would imply $c(x,y) > c(x,y')$ and $c(x',y') > c(x',y)$, two inequalities whose sum contradicts (2.1).

We therefore turn our attention to the ordering (a) $x > x' \geqslant y > y'$ opposite to figure 3*a*. Express

$$x - y = (1-t)(x-y') + t(x'-y)$$

as a convex combination of the larger and smaller quantities $x - y' > x' - y \geqslant 0$. The ordering (a) gives $0 < t < 1$ while summing shows

$$x' - y' = t(x-y') + (1-t)(x'-y).$$

Applying strict concavity of h yields

$$c(x,y) > (1-t)c(x,y') + tc(x',y),$$
$$c(x',y') > tc(x,y') + (1-t)c(x',y),$$

two inequalities whose sum again contradicts (2.1). Therefore, (2.1) must prevent crossings of the circles $O(x,y)$ and $O(x',y')$.

(ii) Assume in addition to (2.1), that the circle $O(x,y)$ is enclosed by $O(x',y')$ and $(y-x)(y'-x') < 0$. The only orderings consistent with these hypotheses are $x' \leqslant y < x \leqslant y'$ or the reverse $x' \geqslant y > x \geqslant y'$ (figure 3*b*). Since x lies halfway between y and $2x-y$ one has $c(x,y) = c(x,2x-y)$. The ordering makes it clear that x' is closer to y than to $2x-y$, so strict monotonicity of h yields $c(x',y) < c(x',2x-y)$. From (2.1),

$$c(x,2x-y) + c(x',y') < c(x,y') + c(x',2x-y). \tag{2.2}$$

The no-crossing rule (i) then implies that the circle $O(x,2x-y)$ does not touch $O(x',y')$, precluding figure 3*b*; one cannot have $2x-y = y'$ since (2.2) is strict. Thus, both x and $2x-y$ lie strictly between x' and y'. On the other hand, the hypotheses are symmetrical in xs and ys, so $2y-x$ also lies between y' and x'. One therefore concludes that the circle $O(2x-y, 2y-x)$ is strictly enclosed by $O(x',y')$. ■

Since the no-crossing rule plays a central role throughout the paper, we extend consideration to all costs for which this rule is satisfied, by introducing the *costs of concave type*. A differential characterization of these costs is given in Appendix B.

Definition 2.2. A function $c : \mathbb{R}^2 \longrightarrow \mathbb{R} \cup \{-\infty\}$ is said to be of *concave type* if inequality (2.1) implies that the circles $O(x,y)$ and $O(x',y')$ do not intersect unless $x = x'$ or $y = y'$.

After two more definitions, we state a theorem which combines our lemma with Monge's observation.

Definition 2.3. A subset $\Omega \subset \mathbb{R}^2$ of the plane *has no crossings* if $(x,y),(x',y') \in \Omega$ implies the circles $O(x,y)$ and $O(x',y')$ do not intersect except perhaps tangentially. If the same hypotheses yield conclusion (i) of lemma 2.1, we say Ω has the *strict no-crossing property*; if they yield conclusion (ii), then Ω *satisfies the rule of three*.

If a set $\Omega \subset \mathbb{R}^2$ has no crossings, then neither will its reflection through the origin or through the line $x = y$. Nor will its closure $\bar{\Omega}$: one cannot obtain a pair of circles that intersect non-tangentially as a limit of a sequence of circles $O(x_n, y_n)$ and $O(x'_n, y'_n)$ that do not.

Definition 2.4. A measure $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ may be said to have *property X* (e.g. have no crossings, satisfy the rule of three), if $\operatorname{spt}\gamma \subset \mathbb{R}^2$ has *property X*.

Theorem 2.5. *For continuous costs $c \geqslant 0$ of concave type, optimal measures $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ have the strict no-crossing property. If $c(x, y) = h(|x - y|)$, they satisfy the rule of three.*

Proof. Let $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ be optimal. For (x, y) and (x', y') from $\operatorname{spt}\gamma$, (2.1) is a particular consequence of Smith & Knott's (1992) characterization of optimal measures, stated here as theorem 5.3. The assumption that the cost be of concave type allows one to conclude that $O(x, y)$ and $O(x', y')$ do not intersect unless $x = x'$ or $y = y'$. Thus, the measure γ has the strict no-crossing property. For costs $c(x, y) = h(|x-y|)$ of concave type, lemma B 4 (in Appendix B) guarantees h to be strictly concave increasing. Applying lemma 2.1(ii) then yields the conclusion that γ satisfies the rule of three. ■

In view of this theorem, the balance of our efforts here will be devoted to analysing measures γ with the strict no-crossing property†. To begin, one should observe that if $x \neq y$ and $y \neq z$, one cannot have both (x, y) and (y, z) in $\operatorname{spt}\gamma$: the circles $O(x, y)$ and $O(y, z)$ will intersect. Thus for $\gamma \in \Gamma(\mu, \nu)$, the strict no-crossing property implies that any mass common to μ and ν will be concentrated along the diagonal $D := \{(x, x) \mid x \in \mathbb{R}\}$ by the measure γ (see, for example, Gangbo & McCann (1996, proof of prop. 2.9)). For optimal γ this comes as no surprise: the costs of concave type satisfy a strict triangle inequality (B 1), so one's first choice should always be to supply a factory from an on-site mine. Any geographically overlapping production and consumption can therefore be subtracted from the problem *a priori*, and one assumes μ and ν mutually singular without loss. It is then convenient to encode the distribution of mines and factories as a single measure $\rho = \mu - \nu$, from which $\mu = \rho_+$ and $\nu = \rho_-$ can be recovered as its positive and negative parts. This signed measure ρ represents excess production; it lies in the space $\mathcal{M}_0(\mathbb{R})$ of neutral measures $\rho[\mathbb{R}] = 0$ with finite total variation.

To understand the structure of sets and measures without crossings, it is helpful to convert the real line into a circle $\mathring{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ by adding a point at infinity (joining $+\infty$ to $-\infty$). Any three distinct points $x_1, x_2, x_3 \in \mathbb{R}$ traverse this circle $\boldsymbol{S}^1 \approx \mathring{\mathbb{R}}$ in a definite direction—either clockwise or counterclockwise—depending on the sign of $(x_1 - x_2)(x_2 - x_3)(x_3 - x_1)$. A key property of sets without crossings turns out to be that they decompose into orientation reversing components.

Definition 2.6. A subset $\Omega \subset \mathbb{R}^2$ of the plane (or of the torus $\mathring{\mathbb{R}} \times \mathring{\mathbb{R}}$) is *orientation reversing* if the triples x_1, x_2, x_3 and y_1, y_2, y_3 traverse the circle $\boldsymbol{S}^1 \approx \mathring{\mathbb{R}}$ in opposite directions whenever $(x_i, y_i) \in \Omega$ for $i = 1, 2, 3$. Here opposite directions mean

$$(x_1 - x_2)(x_2 - x_3)(x_3 - x_1)(y_1 - y_2)(y_2 - y_3)(y_3 - y_1) \leqslant 0. \tag{2.3}$$

† The support of one such measure is depicted by the solid lines in figure 1. Being optimal, this measure also satisfies the rule of three.

Example 2.7. The function $H(x) = 1/x$ extends to an orientation reversing homeomorphism of the circle $\mathring{\mathbb{R}}$ which fixes the points $x = \pm 1$. Its graph $(x, 1/x)$ is an orientation reversing subset of the plane.

If the set without crossings lies in a product $I \times J$ of disjoint intervals, then it consists of a single orientation reversing component. Indeed, $\Omega \subset \mathbb{R}^2$ together with its reflection $\Omega^\dagger$ in the line $y = x$ will be orientation reversing. This follows from the next lemma, which also characterizes this geometry in terms of monotonicity.

Definition 2.8. A set $\Omega \subset \mathbb{R}^2$ is *non-decreasing* if (x, y) and (x', y') in Ω imply $(x' - x)(y' - y) \geqslant 0$.

The lemma is easiest to see by transforming our perspective (i.e. figures 2 and 3) from the upper half-plane $\boldsymbol{H} = \{z = x + \mathrm{i}y \mid y \geqslant 0\}$ to the unit disk $\boldsymbol{D} = \{|w| \leqslant 1\}$. This may be accomplished, for example, by a change of variables in the complex plane:

$$w(z) = \frac{\mathrm{i} - z}{\mathrm{i} + z}, \tag{2.4}$$

the Möbius transformation sending $(0, 1, \infty)$ to $(1, \mathrm{i}, -1)$. The homeomorphism $w : \boldsymbol{H} \longrightarrow \boldsymbol{D}$ preserves angles and circles (up to the boundary where w maps the real line onto the unit circle $\partial\boldsymbol{D}$). Thus two circles cross in the upper half-plane precisely when the image circles cross in the unit disk. Hereafter, w is tacitly used to switch between these two models: we tend to say $x, y \in \mathbb{R}$ when we are thinking about the upper half-plane, as opposed to $x, y \in \mathring{\mathbb{R}}$ or $\in \boldsymbol{S}^1$ when thinking about the disk. By *interval* we always mean a connected subset I of a one-dimensional manifold, though this highlights a difference between the circle and the line: the complement of I in $\mathring{\mathbb{R}}$ will also be an interval, though its complement in $\mathbb{R}$ need not. After the following lemma, our intervals will typically lie on the circle $\mathring{\mathbb{R}}$ (as in figure 4*a*).

Lemma 2.9. *Let $I \subset \mathbb{R}$ be an interval and $J := \mathbb{R} \setminus I$ its complement. Suppose $H : \mathring{\mathbb{R}} \longrightarrow \mathring{\mathbb{R}}$ is an orientation reversing homeomorphism of the circle that fixes both end-points of I. For $\Omega \subset I \times J$, the following conditions are equivalent:*

(A) *Ω has no crossings;*

(B) *$\Omega \cup \Omega^\dagger$ is orientation reversing, where $\Omega^\dagger := \{(x, y) \mid (y, x) \in \Omega\}$; and*

(C) *$\{(x, H(y)) \mid (x, y) \in \Omega\}$ is a non-decreasing subset of the plane.*

Proof of lemma 2.9. Assume that neither I nor J consists of a single point; otherwise (A) $I \times J$ has no crossings, (B) $(I \times J) \cup (J \times I)$ is orientation reversing and (C) $I \times H(J)$ is non-decreasing, so the lemma follows trivially. Fix a set $\Omega \subset I \times J$ in the plane. We shall show (A) $\Rightarrow$ (C) $\Rightarrow$ (B) $\Rightarrow$ (A).

(A) $\Rightarrow$ (C). To show the contrapositive, assume (C) fails. Then there exist points (x_1, y_1) and (x_2, y_2) in $\Omega \subset I \times J$ satisfying

$$(x_2 - x_1)(H(y_2) - H(y_1)) < 0. \tag{2.5}$$

Interchanging $(x_1, y_1) \leftrightarrow (x_2, y_2)$, if necessary, yields $x_1 < x_2$ and $H(y_2) < H(y_1)$. Here $y_1, y_2 \in J = \mathbb{R} \setminus I$. Since H is an orientation reversing homeomorphism of the circle, it swaps the interiors of I and $\mathring{\mathbb{R}} \setminus I$. Thus, all four points $x_1, x_2, H(y_1), H(y_2) \in$

$\mathbb{R}$ lie in the closure of $I \subset \mathbb{R}$. It follows that the triples x_1, x_2, ∞ and $H(y_2), H(y_1), \infty$ are both ordered counterclockwise on the circle. Since the y_i lie outside $I \supset \{x_1, x_2\}$ while H reverses orientations, x_1, x_2, y_1, y_2 are distinct and ordered counterclockwise around the circle. This shows that $O(x_1, y_1)$ intersects $O(x_2, y_2)$ non-tangentially, precluding the possibility (A) that Ω has no crossings and establishing the first implication.

(C) $\Rightarrow$ (B). Assume (C), and choose three points (x_1, y_1), (x_2, y_2) and (x_3, y_3) from $\Omega \cup \Omega^\dagger$. Take the x_i to be distinct, since otherwise (2.3) holds trivially, and similarly, with the y_i. Of the six coordinates $x_1, x_2, x_3, y_1, y_2, y_3$, exactly three must lie in I while the other three lie in J. Henceforth, we shall suppose that in I the x_i are outnumbered by y_i either (i) zero to three; or (ii) one to two; (the other two cases are handled by the same argument but with xs replacing ys). Renumbering if necessary, in case (i) we assume $y_1 < y_2 < y_3$, while in case (ii) we assume $y_1 < y_2$ in I. Since $(y_1, x_1), (y_2, x_2) \in \Omega$, it follows from (C) and $y_1 < y_2$ that $H(x_1) \leqslant H(x_2)$. Similarly, (C) implies $H(x_2) \leqslant H(x_3)$ in case (i). In case (ii) on the other hand, we know that $y_1 < y_2$ and $H(x_1) \leqslant H(x_2)$ both lie in the closure of $I \subset \mathbb{R}$, while y_3 and $H(x_3)$ lie in the closure of $\mathbb{R} \setminus I$. In either case, both y_1, y_2, y_3 and $H(x_1), H(x_2), H(x_3)$ are ordered counterclockwise around the circle. Since H reverses orientations, x_1, x_2, x_3 must be ordered clockwise. This establishes (B) that $\Omega \cup \Omega^\dagger$ is orientation reversing, hence the second implication is proved.

(B) $\Rightarrow$ (A). Finally, assume that $\Omega \cup \Omega^\dagger$ is orientation reversing and select (x_1, y_1) and (x_2, y_2) from Ω. If the circles $O(x_1, y_1)$ and $O(x_2, y_2)$ were to intersect non-tangentially, then the four points x_1, x_2, y_1, y_2 would be distinct and ordered either clockwise or counterclockwise around the circle. Defining $(x_3, y_3) := (y_1, x_1) \in \Omega^\dagger$, this would contradict the assumption that $x_1, x_2, x_3 = y_1$ and $y_1, y_2, y_3 = x_1$ traverse the circle in opposite directions. Thus (A) Ω has no crossings. ■

Now suppose the density of $\rho \in \mathcal{M}_0(\mathbb{R})$ changes sign only twice along the line. A characterization for the measures $\gamma \in \Gamma(\rho_+, \rho_-)$ with no crossings follows immediately.

Corollary 2.10. *Given an interval $I \subset \mathring{\mathbb{R}}$ with complement $J := \mathring{\mathbb{R}} \setminus I$, suppose the measure $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ assigns full mass to $I \times J$. Define its reflection $\gamma^\dagger$ by $\gamma^\dagger[\Omega] := \gamma[\Omega^\dagger]$ for Borel $\Omega^\dagger = \{(y, x) \in \mathbb{R}^2 \mid (x, y) \in \Omega\}$. Then γ has no crossings if and only if $\gamma + \gamma^\dagger$ is orientation reversing.*

Proof of corollary 2.10. The set $\Omega := \operatorname{spt}\gamma \cap (I \times J)$ has full mass for γ, while $\Omega \cup \Omega^\dagger$ has full mass for $\gamma + \gamma^\dagger$. If γ has no crossings, then neither will $\Omega \subset \operatorname{spt}\gamma$; lemma 2.9 implies that $\Omega \cup \Omega^\dagger$ is orientation reversing, as is its closure $\operatorname{spt}[\gamma + \gamma^\dagger]$ in view of (2.3). Conversely, if $\Omega \cup \Omega^\dagger \subset \operatorname{spt}[\gamma + \gamma^\dagger]$ is orientation reversing, the same lemma shows Ω will have no crossings. Neither will its closure $\operatorname{spt}\gamma$. ■

Combined with theorem 2.5, this characterization makes it clear that if $\mu - \nu \in \mathcal{M}_0(\mathbb{R})$ changes signs only twice along the line, then the optimal measure γ in $\Gamma(\mu, \nu)$ has orientation reversing support. This is analogous but exactly opposite to the result for strictly convex costs, where optimality selects the unique measure with non-decreasing support. By contrast, when the supports of μ and ν lie on opposite sides of some point a, then the reflection $H(x) = 2a - x$ in lemma 2.9 shows $\operatorname{spt}\gamma$ is non-*in*creasing. Nonetheless, the orientation reversing geometry provides enough rigidity

to select γ uniquely. This is the content of the next proposition. It is derived from the convex case using a standard fact of measure theory: any map $Z : \boldsymbol{X} \longrightarrow \boldsymbol{Y}$ on a measure space $(\boldsymbol{X}, \lambda)$ induces an image measure $Z_{\#}\lambda$ on $\boldsymbol{Y}$, defined for $U \subset \boldsymbol{Y}$ by

$$Z_{\#}\lambda[U] := \lambda[Z^{-1}(U)]; \tag{2.6}$$

here Z and U are assumed measurable with respect to given σ-algebras in $\boldsymbol{X}$ and $\boldsymbol{Y}$. A measurable function $f : \boldsymbol{Y} \longrightarrow \mathbb{R}$ can be integrated against this image measure using the change of variables formula

$$\int_{\boldsymbol{Y}} f(y)\,\mathrm{d}Z_{\#}\lambda(y) = \int_{\boldsymbol{X}} f(Z(x))\,\mathrm{d}\lambda(x). \tag{2.7}$$

Proposition 2.11. *Fix measures $\mu, \nu \in \mathcal{M}_+(\mathbb{R})$. Suppose ν vanishes on some interval $I \subset \mathring{\mathbb{R}}$ while μ vanishes on its complement $J := \mathring{\mathbb{R}} \setminus I$. Then only one joint measure, $\gamma \in \Gamma(\mu, \nu)$, has no crossings.*

Proof of proposition 2.11. Assume some $\gamma \in \Gamma(\mu, \nu)$ has no crossings, since otherwise the proposition holds vacuously; in particular, μ and ν must have the same total mass $\gamma[\mathbb{R}^2]$. We need to show that γ is specified uniquely by μ and ν.

Interchange μ with ν if necessary to assume $\infty \in J$. Choose an orientation reversing homeomorphism $H : \mathring{\mathbb{R}} \longrightarrow \mathring{\mathbb{R}}$ of the circle that fixes both end-points of I; for example, $H(y) = z + r^2/(y - z)$ works nicely when these end-points $z \pm r$ are finite and distinct. Extend H to a homeomorphism $Z(x, y) = (x, H(y))$ of the torus $\mathring{\mathbb{R}} \times \mathring{\mathbb{R}}$. This homeomorphism induces a bijection, $\gamma \leftrightarrow Z_{\#}\gamma$, given by (2.6) between the two collections of measures $\Gamma(\mu, \nu)$ and $\Gamma(\mu, H_{\#}\nu)$.

Now $\Omega = \operatorname{spt}\gamma \cap I \times J$ has no crossings and carries full mass for γ. By lemma 2.9 its image $Z(\Omega)$ is non-decreasing in the plane, as is the closure of this image: $\operatorname{spt} Z_{\#}\gamma \subset \mathbb{R}^2$. Thus $Z_{\#}\gamma$ must be the unique measure with non-decreasing support in $\Gamma(\mu, H_{\#}\nu)$; its uniqueness is well known, though a direct proof is provided in Appendix A by proposition A 2. Since $Z_{\#}$ acts bijectively on $\mathcal{M}_0(\mathbb{R})$, γ too is uniquely determined by μ and ν. ∎

Thus when $\mu - \nu \in \mathcal{M}_0(\mathbb{R})$ changes signs only twice along the line, the optimal measure γ in $\Gamma(\mu, \nu)$ is cost-independent among all costs of concave type: it is determined uniquely by the requirement that its support has no crossings. In the next section, we proceed to analyse the complications that arise when $\mu - \nu$ oscillates more than twice along the line. For this purpose, it is useful to have an explicit formula for the unique measure of proposition 2.11.

Recall that any measure $\mu \in \mathcal{M}_+(\mathbb{R})$ on the line can be represented by a non-decreasing function $X : [0, t] \longrightarrow \mathbb{R} \cup \{\pm\infty\}$; here $t := \mu[\mathbb{R}]$ denotes total mass, while X can be defined by $X(\theta) := \sup\{x \mid \mu[(-\infty, x)] < \theta\}$. Then $\mu = X_{\#}\lambda$ is recovered by pushing the Lebesgue measure λ forward through the mapping $\theta \longrightarrow X(\theta)$ using (2.6). If there is ambiguity about the domain of $X(\theta)$, the restriction of Lebesgue to the interval $[0, t]$ may be denoted $\lambda_{[0,t]}$.

For intervals $I \subset \mathbb{R}$, we adopt the notation $(-\infty, I) \subset \mathbb{R}$ to denote any connected component of $\mathbb{R} \setminus I$ that extends to infinity in the negative direction.

Proposition 2.12. *Let $\mu, \nu \in \mathcal{M}_0(\mathbb{R})$ have the same total mass $t = \mu[\mathbb{R}]$. Assume ν vanishes on some interval $I \subset \mathring{\mathbb{R}}$ while μ vanishes on its complement $J := \mathring{\mathbb{R}} \setminus I$. Represent $\mu = X_{\#}\lambda_{[0,t]}$ and $\nu = Y_{\#}\lambda_{[0,t]}$ by non-decreasing X and $Y : [0, t] \longrightarrow \mathbb{R}$,*

and extend $Y(\theta) = Y(\theta + t)$ *periodically. Set* $\phi := \nu[(-\infty, I)]$ *if* $I \subset \mathbb{R}$, *and* $\phi := \mu[(-\infty, J)]$ *otherwise. Then the curve* $Z(\theta) := (X(\theta), Y(\phi - \theta))$ *on* $0 < \theta < t$, *supports a measure* $\gamma := Z_{\#}\lambda_{[0,t]}$ *in* $\Gamma(\mu, \nu)$ *with no crossings.*

Proof of proposition 2.12. First define an orientation reversing homeomorphism $H : \mathring{\mathbb{R}} \longrightarrow \mathring{\mathbb{R}}$ that fixes the end-points of I, and assume that $\infty \in J$. Then ν has mass ϕ to the left of $I \subset \mathbb{R}$ and mass $t - \phi$ to its right, so $Y(\phi - \theta)$ moves clockwise from one end of J to the other as θ increases from 0 to t. Because H is orientation reversing and exchanges the interiors of J and I, it follows that $H(Y(\phi - \theta))$ is non-decreasing on $0 < \theta < t$. Thus the curve $(X(\theta), H(Y(\phi - \theta)))$ is non-decreasing in the plane. By lemma 2.9, the original curve $Z(\theta)$ in $I \times J$ will not have crossings. This curve contains the full mass of $Z_{\#}\gamma$, so neither its closure nor $\mathrm{spt}[Z_{\#}\gamma]$ can have crossings. Since $Z_{\#}\gamma$ has marginals $X_{\#}\lambda_{[0,t]} = \mu$ and $Y_{\#}\lambda_{[0,t]} = \nu$, it must be the (unique) measure in $\Gamma(\mu, \nu)$ with no crossings.

Now suppose $\infty \in I$. Extending $X(\theta) = X(\theta + t)$ periodically, it is shown above that the curve $Z^{\dagger}(\theta) := (Y(\theta), X(\phi - \theta))$ on $0 < \theta < t$ has no crossings. Noting that $Z^{\dagger}(\phi - \theta) \in J \times I$ represents the reflection of $Z(\theta)$ in the line $x = y$, it follows that neither $Z(\theta)$ nor $\gamma = Z_{\#}\lambda_{[\phi - t, \phi]} \in \Gamma(\mu, \nu)$ will have crossings. ■

Remark 2.13. Let $I \subset \mathring{\mathbb{R}}$ be an interval with end-points a and b. A restatement of the same result asserts that if $X(\theta)$ moves counterclockwise through I while $Y(\theta)$ moves clockwise through $\mathring{\mathbb{R}} \setminus I$ from $X(0) = Y(0) = a$ to $X(t) = Y(t) = b$, then the curve $Z(\theta) = (X(\theta), Y(\theta))$ on $0 < \theta < t$ and measure $Z_{\#}\lambda_{[0,t]}$ have no crossings.

3. Networks and the transportation hierarchy

This section is devoted to developing a change of variables that simplifies the transportation problem when the density of $\mu - \nu$ changes sign finitely often, say $2m + 2$ times, along the line. For $m = 0$, we have already seen that the no-crossing rule selects a unique measure $\gamma \in \Gamma(\mu, \nu)$, simultaneously optimal for all costs of concave type. When $m > 0$, it is no longer true that only one measure in $\Gamma(\mu, \nu)$ has no crossings. What does remain true is that measures without crossings form a finite-dimensional subset of the infinite-dimensional space $\Gamma(\mu, \nu)$. In the new variables, this set can be described as a finite union of convex polytopes, each having dimension no greater than m. There are $(3m \text{ choose } m)/(2m + 1)$ of these polytopes, corresponding to different choices for which intervals of positive mass will supply locally with respect to the others, i.e. to what one might call the local-global hierarchy among intervals of supply. In the present section, we develop this finite-dimensional parametrization of measures without crossings, summarized by the bijection of theorem 3.11. Two ingredients that enter the discussion will be the language of networks (i.e. directed graphs) and the duality theory for planar graphs summarized in Rockafellar (1984).

Given a signed measure $\rho \in \mathcal{M}_0(\mathbb{R})$, an *interval of supply* refers to a maximal interval $I \subset \mathring{\mathbb{R}}$ on which ρ_- vanishes while $\rho_+[I] > 0$. We use the notation $\rho \in \mathcal{M}_0^m(\mathbb{R})$ to indicate that the full mass of ρ_+ is contained in $m + 1$ intervals of supply, in which case we say that ρ changes sign $2m + 2$ times along the line; here $m \geqslant 0$ is an integer. Maximality ensures that intervals of supply must be disjoint. For $\rho \in \mathcal{M}_0^m(\mathbb{R})$, these intervals $I_0, \ldots, I_m$ can be numbered counterclockwise around the circle. The complement $\mathring{\mathbb{R}} \setminus \bigcup I_j$ also consists of $m + 1$ intervals $J_0, \ldots, J_m$, on which ρ_+ vanishes though ρ_- does not, to be called *intervals of demand.*

Now suppose that $\gamma \in \Gamma(\rho_+, \rho_-)$ is a competitor in the problem of transporting ρ_+ onto ρ_-. Then the full mass of γ lies in $\bigcup_{i,j=0}^{m} I_i \times J_j$. Moreover, if γ is optimal and $m > 1$, the no-crossing rule of the preceding section will force γ to vanish on some of the rectangles $I_i \times J_j$. To keep track of such implications, it is helpful to introduce the following graph-theoretic structure on the intervals of supply and demand, or equivalently (and hereafter tacitly) on $2m+2$ representative points chosen arbitrarily from these intervals, so that $x_{2i+1} \in I_i$ and $x_{2i+2} \in J_i$ for $i = 0, \ldots, m$.

Definition 3.1. An *uncrossed network* refers to a finite graph G satisfying (i)–(iv). Its vertices and edges are denoted by $\mathrm{node}(G)$ and $\mathrm{arc}(G) \subset \mathrm{node}(G) \times \mathrm{node}(G)$.

(i) The vertices of G consist of an even number of distinct points on the unit circle: $\mathrm{node}(G) = \{x_1, \ldots, x_{2m+2}\} \subset \mathring{\mathbb{R}}$, enumerated counterclockwise.

(ii) If $(x_i, x_j) \in \mathrm{arc}(G)$, then i must be odd and j must be even.

(iii) $\mathrm{arc}(G) \subset \mathring{\mathbb{R}}^2$ has no crossings.

(iv) No further arcs (x_i, x_j) could be added to $\mathrm{arc}(G)$ without violating (ii) or (iii).

The nodes $\{x_1, \ldots, x_{2m+2}\}$ of such a network form the vertices of a polygon P inscribed in the unit circle. Representing the arcs of G by straight lines, from (iv) it follows that these lines include all edges of this polygon and subdivide its interior into smaller convex regions (figure 4*a*). Each of these interior regions F will be a quadrilateral: F must have an even number of vertices since the network is bipartite (ii), and their number cannot exceed four by the maximality condition (iv). There are m quadrilaterals in total. Together with the $2m + 2$ domains lying between the edges of our polygon and the unit circle, they constitute the set $\mathrm{face}(G)$ of faces of G; the quadrilaterals may be referred to as *interior faces* and the remaining domains as *boundary faces*. In the well-known duality theory for planar graphs, the faces of G correspond to nodes of the dual network G^* to G (see figure 4*b*). The arcs $(L, R) \in \mathrm{arc}(G^*)$ are defined to consist of those pairs of faces $L, R \in \mathrm{face}(G)$ that share a side $(I, J) \in \mathrm{arc}(G)$, with L lying to the left of an arrow joining I to J and R lying to its right; thus the arcs of G^* are in a one-to-one correspondence with the arcs of G. Just as the Catalan numbers count divisions of P into triangles, the number d_m of uncrossed networks can be computed recursively. Fixing m and the x_i, the computation is a special case of a result of Erdélyi & Etherington (1940): solving the recursion

$$d_m = \sum_{i=1}^{m} \sum_{j=i}^{m} d_{i-1} d_{j-i} d_{m-j},$$

with $d_0 = 1$, they find $d_m = (3m \text{ choose } m)/(2m + 1)$. The $d_2 = 3$ possibilities for $m = 2$ are shown in figure 5.

Functions on the nodes, arcs and faces of an uncrossed network G will be referred to as *densities*, *flows* and *potentials*, respectively. Each flow $s : \mathrm{arc}(G) \longrightarrow \mathbb{R}$ induces a density $\rho : \mathrm{node}(G) \longrightarrow \mathbb{R}$, denoted $\rho = \operatorname{div} s$, which measures the net flux

$$\operatorname{div} s(I) := \sum_{(I,J) \in \mathrm{arc}(G)} s(I, J) - \sum_{(J,I) \in \mathrm{arc}(G)} s(J, I) \tag{3.1}$$

out of $I \in \mathrm{node}(G)$. Similarly, each potential $\phi : \mathrm{face}(G) \longrightarrow \mathbb{R}$ induces a flow $s = \operatorname{curl} \phi$ on $\mathrm{arc}(G)$, defined by $\operatorname{curl} \phi(a) := \phi(R) - \phi(L)$, where $L, R \in \mathrm{face}(G)$ denote the faces lying immediately to the left and right of $a \in \mathrm{arc}(G)$ (see figure 6).

Definition 3.2. Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and an uncrossed network G on its intervals of supply and demand. A measure $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ is said to be *compatible with* G if γ has no crossings and $\gamma[I \times J] = 0$ whenever $(I, J) \in \mathrm{node}(G)^2$ is not an arc of G.

The set of measures γ compatible with G and having marginals μ and ν will be denoted by $\Gamma_G(\mu, \nu)$.

Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and an uncrossed network G on its intervals of supply and demand. In the next lemma and following, we tacitly identify the nodes of G with intervals $I \subset \mathring{\mathbb{R}}$ of supply or demand, while identifying the measure $\rho \in \mathcal{M}_0^m(\mathbb{R})$ with the density $\rho(I) := \rho[I]$ that it induces on $I \in \mathrm{node}(G)$.

Lemma 3.3. *Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$. If $\gamma \in \Gamma(\rho_+, \rho_-)$ has no crossings, then γ is compatible with some uncrossed network G on ρ's intervals of supply and demand.*

Proof. Choose $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and $\gamma \in \Gamma(\rho_+, \rho_-)$. Let $\mathrm{node}(G)$ denote the intervals of supply and demand for ρ (or to be pedantic choose representative points $x_i \in \boldsymbol{S}^1$ from each of these $2m + 2$ intervals). Include in $\mathrm{arc}(G)$ all pairs of intervals $(I, J) \in \mathrm{node}(G) \times \mathrm{node}(G)$ with $\gamma[I \times J] > 0$. For an interval I of supply, $\gamma[\mathbb{R} \times I] = \rho_-[I]$ vanishes by definition; moreover, the intervals of supply for $\rho \in \mathcal{M}_0^m(\mathbb{R})$ are assumed to contain the full mass of ρ_+, so $\gamma[J \times \mathbb{R}]$ vanishes whenever J is an interval of demand. Thus $(I, J) \in \mathrm{arc}(G)$ forces I to be an interval of supply and J to be an interval of demand. Since intervals of supply alternate with intervals of demand around the circle $\boldsymbol{S}^1$, conditions (i) and (ii) of definition 3.1 follow easily.

To verify (iii), suppose (I_0, J_0) and $(I_1, J_1) \in \mathrm{arc}(G)$ and denote their representative points by $x_i \in I_i$ and $y_i \in J_i$. We need to know that the circle $O(x_0, y_0)$ does not cross $O(x_1, y_1) \subset \mathbb{R}^2$; to derive a contradiction, assume these circles intersect non-tangentially. Then the four points x_0, x_1, y_0, y_1 must be distinct, and ordered either clockwise or counterclockwise around the circle; suppose clockwise without loss. The intervals I_0, I_1, J_0, J_1 that they represent must also be distinct—therefore disjoint—and occur in the same order on the circle. Since $\gamma[I_i \times J_i] > 0$, there exist $(x_i', y_i') \in (I_i \times J_i) \cap \mathrm{spt}\,\gamma$ for $i = 0, 1$. The points x_0', x_1', y_0', y_1' come from disjoint intervals in clockwise order on the circle, so $O(x_0', y_0')$ intersects $O(x_1', y_1')$ non-tangentially. But this contradicts the assumption that γ has no crossings. Thus we conclude (iii) that $O(x_0, y_0)$ does not cross $O(x_1, y_1)$.

Although the maximality condition (iv) may not yet be satisfied, we can add additional pairs (I, J) from $\mathrm{node}(G) \times \mathrm{node}(G)$ to $\mathrm{arc}(G)$, as long as the addition of each new pair does not violate (ii) or (iii). By finiteness of $\mathrm{node}(G)$, this process terminates. The construction ensures compatibility of γ with the resulting uncrossed network G. ■

In a network G, a *path* refers to a sequence $I_1, I_2, \ldots, I_n \in \mathrm{node}(G)$, each node linked to the next either by a forward arc, $(I_{i-1}, I_i) \in \mathrm{arc}(G)$, or a backward arc, $(I_i, I_{i-1}) \in \mathrm{arc}(G)$, $i = 2, \ldots, n$. When $I_n = I_1$, the path is said to be a *circuit*, while if a path uses each arc at most once—either forward or backward, but not both—it is said to be *elementary*. For the network G to be *path connected* means that any two nodes can be joined by a path. A path-connected network with no elementary circuits is known as a *tree*.

The next lemma asserts that the dual network G^* to an uncrossed network G will be a tree (an example is shown in figure 4*b*). Since the interior faces of G are quadrilaterals with supply nodes and demand nodes at diagonally opposite vertices,

each interior node, $F \in \mathrm{node}(G^*)$, of the dual graph has two incoming arcs and two outgoing arcs connecting it to four adjacent nodes in G^*. The boundary nodes of G^* are connected to a single adjacent node by one arc, either incoming or outgoing. These boundary nodes are ordered around the circle $\mathbb{R} \approx \boldsymbol{S}^1$; such a boundary node $B \in \mathrm{node}(G^*)$ is said to be a *root* of G^* if B and both of its neighbours on the circle are all connected to the same interior node (by single forward or backward arcs); thus F_4 and F_7 form the two roots in figure 4*b*.

Lemma 3.4. *The planar dual G^* of an uncrossed network G will be a tree. If G^* has more than two nodes, then at least two of them are roots.*

Proof. Recall that G can be represented as a convex polygon P inscribed in the unit disk $\boldsymbol{D} \subset \mathbb{R}^2$, and subdivided into m quadrilaterals sharing their vertices with P (see figure 4). These quadrilaterals represent the interior nodes of G^*, while the $2m+2$ components of the complement $\boldsymbol{D} \setminus P$ correspond to boundary nodes of G^*. The proof that G^* is a tree will proceed by induction on m.

When $m = 0$, the polygon P degenerates to a line segment and G^* consists of two (boundary) nodes connected by a single arc: so G^* is a tree. Now suppose the lemma has also been established for all smaller values of $m > 0$. Choose an interior face $Q \in \mathrm{node}(G^*)$, such as the face F_2 of figure 4*a*. The quadrilateral Q has its vertices on the unit circle. If the interior of Q is removed from the polygon P, it leaves behind four smaller (or possibly degenerate) convex polygons, each inscribed in the unit circle and subdivided into a total of $m-1$ quadrilaterals. The four corresponding networks G_i are uncrossed, $i = 1, 2, 3, 4$, while their duals G_i^* are disjoint except that Q acts as a boundary node for each of them. By the inductive hypothesis, each G_i^* is a (path-connected) tree. Thus there is a path from $Q \in \mathrm{node}(G_i^*)$ to any other node in G_i^*, so $\mathrm{node}(G^*) = \bigcup_{i=1}^4 \mathrm{node}(G_i^*)$ must also be path connected. If G^* failed to be a tree, that would imply the existence of an elementary circuit not contained entirely in any of the trees G_i^*. Such a circuit must include Q as a node, followed by some adjacent node, say $F \in \mathrm{node}(G_1^*)$. Since the circuit must leave and re-enter G_1^* through the arc (F, Q) or (Q, F), it uses the same arc both forwards and backwards. But this violates the definition of an elementary circuit, proving that G^* is a tree.

To show that G^* has two roots, consider three cases: either (i) all four sub-networks G_i^* consist of a single arc; or (ii) three of the four sub-networks consist of a single arc; or (iii) at least two of the G_i^* have more than one arc. In the first case, Q is connected directly to all four boundary nodes of G^*, each of which is a root of G^*. Turning to case (ii), three of the arcs that begin or end at Q are connected directly to boundary nodes; the middle of these is a root of G^*. The fourth arc connects Q to a sub-network G_4^* consisting of more than two nodes. But the inductive hypothesis implies that G_4^* contains at least two roots, which may also be roots of G^* depending on their positions relative to Q. Let $L, R \in \mathrm{node}(G_4^*)$ denote the boundary nodes of G_4^* to the immediate left and right of Q. Apart from Q, all boundary nodes of G_4^* are also boundary nodes of G^*. Thus any root of G_4^*, save for L, Q or R, will also be a root of G^*. If G_4^* has only one root among L, Q and R, then its second root provides a second root for G^*. If two roots of G_4^* occur among L, Q and R, then $m = 2$ and the fourth boundary node of G_4^* will be a root of G^* as well as of G_4^*. This concludes case (ii). In the final case (iii), at least two sub-networks—say G_3^* and G_4^*—have two roots each by the inductive hypothesis. The above argument shows that both G_3^* and G_4^* contribute a root to G^*, thus concluding the proof of the lemma. ■

Since only one elementary path connects each pair of nodes in a tree, this lemma yields the following corollary.

Corollary 3.5. *Every flow s on an uncrossed network G can be expressed as $s = \operatorname{curl}\phi$ for some $\phi : \operatorname{face}(G) \longrightarrow \mathbb{R}$. This potential ϕ is unique up to additive constant. It satisfies $(\operatorname{div} s)(I) = \phi(B) - \phi(A)$ when $A, B \in \operatorname{face}(G)$ are the boundary faces adjacent to $I \in \operatorname{node}(G)$, with A, I, B ordered counterclockwise.*

Proof. Since there is a one-to-one correspondence between arcs of G^* and of G, $s : \operatorname{arc}(G) \longrightarrow \mathbb{R}$ can be viewed equally well as a flow on the tree G^*. As there are no non-zero *circulations* in a tree, Rockafellar (1984, §§ 4F and 1I) asserts that every flow $s : \operatorname{arc}(G^*) \longrightarrow \mathbb{R}$ can be expressed as the *gradient* of a potential $\phi : \operatorname{node}(G^*) \longrightarrow \mathbb{R}$,

$$s(L, R) = \phi(R) - \phi(L), \tag{3.2}$$

for $(L, R) \in \operatorname{arc}(G^*)$. Indeed, we can define $\phi(F)$ at any $F \in \operatorname{node}(G^*)$ by summing the flow s over a path from L to F. The terms in this sum are signed according to whether each arc is traversed forward or backwards, so $\phi(F) - \phi(L)$ is well defined precisely because G^* is a tree: there is only one elementary path between L and F.

Now the faces $L, R \in \operatorname{face}(G)$ lie to the left and right of the corresponding arc (I, J) of G as depicted in figure 6, so $s(I, J) = \operatorname{curl}\phi(I, J)$ follows from (3.2), the identification $\operatorname{face}(G) = \operatorname{node}(G^*)$ and our definition of $\operatorname{curl}\phi$. For the zero flow $s = 0$ in (3.2), connectedness of G^* forces ϕ to be a constant throughout G^*. For more general flows, linearity of curl guarantees uniqueness up to constant of the corresponding ϕ.

Finally, let $A, B \in \operatorname{face}(G)$ be the boundary faces adjacent to $I \in \operatorname{node}(G)$. Now $(\operatorname{div} s)(I)$ measures the net flow out of the node I, or equivalently the net flux crossing the path through G^* from A to B: namely $\phi(B) - \phi(A)$. ■

A further uniqueness property of measures without crossings is derived from proposition 2.11.

Lemma 3.6. *Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and an uncrossed network G on its intervals of supply and demand. Let $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ be compatible with G, and denote its marginals by μ and ν. If the intervals $I, J \in \operatorname{node}(G)$ share an end-point, then the restriction $\gamma|_{I\times J}$ is uniquely determined by μ, ν and the total flow, $t := \gamma[I \times J]$, from I to J.*

Proof. Let a denote the common end-point of I and J, and suppose without loss that $t > 0$ while I, a and J are ordered clockwise around the circle. Let $X(\theta)$ map $\theta \in [0, \mu(I)]$ into the interval I of supply, pushing the Lebesgue measure forward to $\mu|_I = X_{\#}\lambda_{[0,\mu(I)]}$; this map is determined uniquely a.e. by assuming $X(\theta) \in \mathring{\mathbb{R}}$ to move *counterclockwise* from one end of I to the other. Similarly, let $Y : [0, \nu(J)] \longrightarrow J$ be the *clockwise* map pushing the Lebesgue measure forward to the restriction $\nu|_J$ of ν to J. Remark 2.13 shows that neither the curve $Z(\theta) = (X(\theta), Y(\theta))$ on $0 < \theta < t$, nor the measure $Z_{\#}\lambda_{[0,t]}$, will have crossings. The construction of this measure used I, J, μ, ν and t, but not γ. If we can show that the restriction $\gamma^1 := \gamma|_{I\times J}$ of γ to $I \times J$ has the same marginals as $Z_{\#}\lambda_{[0,t]}$, then $\gamma^1 = Z_{\#}\lambda_{[0,t]}$ follows from proposition 2.11. This would prove the lemma.

Define $\gamma^0 = \gamma - \gamma^1$, and denote the marginals of γ^1 by μ^1 and ν^1. We need to show μ^1 and ν^1 coincide with $X_{\#}\lambda_{[0,t]}$ and $Y_{\#}\lambda_{[0,t]}$. It is enough to prove the inequalities $\mu^1 \geqslant X_{\#}\lambda_{[0,t]}$ and $\nu^1 \geqslant Y_{\#}\lambda_{[0,t]}$ since the total masses $\mu^1[\mathbb{R}] = t = \nu^1[\mathbb{R}]$ are the

same. For simplicity, assume initially that μ does not include a point mass at $X(t)$. Then $X_\#\lambda_{[0,t]}$ represents the restriction of μ to the (relatively closed) interval $U \subset I$ whose end-points are a and $X(t)$. Failure of the inequality $\mu^1 \geqslant X_\#\lambda_{[0,t]}$ means that μ^1 does not vanish on $I \setminus U$ while γ^0 does not vanish on $U \times \mathbb{R}$. It would then be possible to find $(x,y) \in (I \setminus U) \times J$ in $\operatorname{spt}\gamma^1$ and $(x',y') \in U \times J'$ in $\operatorname{spt}\gamma^0$, where $J' \neq J \in \operatorname{node}(G)$ is some other interval of demand. Since x, x', y and y' are distinct points ordered clockwise around $\mathring{\mathbb{R}}$, the circles $O(x,y)$ and $O(x',y')$ intersect non-tangentially. This violates compatibility of γ with G, whence $\mu^1 \geqslant X_\#\lambda_{[0,t]}$.

If μ^1 includes a point mass at $X(t)$ the same argument adapts easily: as long as μ^1 does not vanish on $I \setminus U$, no changes are necessary. The only other way for the desired equality to fail is if μ^1 concentrates more mass than $X_\#\lambda_{[0,t]}$ at $X(t)$. But this is also precluded by the preceding argument, provided we redefine $U \subset I$ to be relatively open by excluding the end-point $X(t)$. A similar proof establishes $\nu^1 \geqslant Y_\#\lambda_{[0,t]}$ to conclude the lemma. ■

This lemma fuels an inductive proof, that each measure $\gamma \in \Gamma_G(\rho_+,\rho_-)$ is uniquely determined by the flow it induces on G.

Proposition 3.7. *Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and an uncrossed network G on its intervals of supply and demand. Each measure $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ compatible with G defines a flow $s_\gamma(I,J) := \gamma[I \times J]$ on G satisfying $s_\gamma \geqslant 0$ and $\operatorname{div} s_\gamma = \mu - \nu$. Here the marginals μ and ν of γ are evaluated on* $\operatorname{node}(G)$. *If $\tilde\gamma \in \Gamma_G(\mu,\nu)$ and $s_{\tilde\gamma} = s_\gamma$ then $\tilde\gamma = \gamma$.*

Proof. Let $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ be compatible with G and set $s_\gamma(I,J) := \gamma[I \times J] \geqslant 0$ for $(I,J) \in \operatorname{arc}(G)$. This defines a non-negative flow s_γ on G. Recalling that compatibility means $\gamma[I \times J] = 0$ whenever $(I,J) \in \operatorname{node}(G) \times \operatorname{node}(G)$ is not an arc of G, we recover from (3.1):

$$\begin{aligned}\operatorname{div} s_\gamma(I) &= \sum_{J \in \operatorname{node}(G)} \gamma[I \times J] - \gamma[J \times I] \\ &= \gamma[I \times \mathbb{R}] - \gamma[\mathbb{R} \times I] \\ &= \mu[I] - \nu[I],\end{aligned}$$

where μ and ν are the marginals of γ. The second equality reflects the fact that the circle $\mathring{\mathbb{R}}$ decomposes disjointly into intervals of supply and demand.

The proof that μ, ν and s_γ determine $\gamma \in \Gamma_G(\mu,\nu)$ uniquely will proceed by induction on $m \geqslant 0$. If $m = 0$, then $\operatorname{node}(G)$ decomposes $\mathring{\mathbb{R}}$ into a single interval I of supply and one interval J of demand. Compatibility of γ with G means that γ vanishes outside $I \times J$. Since I and J share an end-point, lemma 3.6 specifies $\gamma = \gamma|_{I \times J}$ uniquely in terms of μ, ν and $s_\gamma(I,J)$.

Therefore, take $m > 0$ and assume that the proposition has been established for all uncrossed networks with fewer than $2m+2$ nodes. Invoking lemma 3.4 yields a boundary face $B \in \operatorname{face}(G)$ corresponding to a root of the dual tree G^*. Now B shares a side $(I,J) \in \operatorname{arc}(G)$ with an interior face $Q \in \operatorname{face}(G)$, whose four vertices may be labelled I, J, I' and J'. The fact that B is a root of G^* simply means that Q shares two other sides, namely (I,J') and $(I',J) \in \operatorname{arc}(G)$, with boundary faces of G. In other words, Q is exposed on three sides like the faces F_1 or F_3 in figure 4*a*, so the intervals J', I, J and $I' \subset \mathring{\mathbb{R}}$ lie adjacent to each other on the circle.

Let $\gamma \in \Gamma(\mu,\nu)$ be compatible with G, and decompose it as $\gamma = \gamma^0 + \gamma^1 + \gamma^2 + \gamma^3$ where γ^1, γ^2 and γ^3 denote the respective restrictions of γ to $I \times J'$, $I \times J$ and $I' \times J$. For $i = 1,2,3$, lemma 3.6 expresses γ^i in terms of μ, ν and s_γ. Denoting the marginals of γ^i by μ^i and ν^i, it follows that $\mu^0 = \mu - \mu^1 - \mu^2 - \mu^3$ and similarly ν^0 are also determined by μ, ν and the flow. Now B was a root, so I and J are not connected to any other nodes of G except through I' and J'. Thus γ^0 is compatible with the sub-network G_0 obtained from G by eliminating the nodes I and J together with all incident arcs. This corresponds to deleting the exposed quadrilateral Q from the polygon P, so the remaining network G_0 will itself be uncrossed. Since $\gamma^0[K \times L] = s_\gamma(K,L)$ for each arc $(K,L) \in G_0$, the inductive hypothesis asserts that γ^0 can be recovered from μ^0, ν^0 and s_γ. Thus,

$$\gamma = \sum_{i=0}^{3} \gamma^i \in \Gamma_G(\mu,\nu)$$

is determined uniquely by s_γ, concluding the proof. ■

Definition 3.8. Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and an uncrossed network, G, on its intervals of supply and demand. A potential $\phi : \mathrm{face}(G) \longrightarrow \mathbb{R}$, is called *feasible* if the flow, $\operatorname{curl}\phi \geqslant 0$, is non-negative and the $2m+2$ boundary conditions, $\phi(B) = \rho[(-\infty, I)]$, are satisfied; here $B \in \mathrm{face}(G)$ denotes the boundary face just clockwise from $I \in \mathrm{node}(G)$.

Remark 3.9. For intervals whose interior lies on the line, the notation $(-\infty, I)$ is used to denote any connected component of $\mathbb{R} \setminus I$ that is unbounded in the negative direction; in the same vein $(-\infty, I] := (-\infty, I) \cup I$. This definition is extended to intervals whose interior includes ∞ by adding the following convention: when $\mathbb{R} \cap I$ is a union of two intervals, one, I_-, unbounded below and the other, I_+, unbounded above, then $(-\infty, I) := \mathbb{R} \setminus I_+$ while $(-\infty, I] := I_-$. For neutral measures $\rho \in \mathcal{M}_0(\mathbb{R})$, these conventions ensure

$$\rho[(-\infty, I]] = \rho[(-\infty, I)] + \rho[I] = \rho[(-\infty, J)], \tag{3.3}$$

when J is a disjoint interval lying adjacent to I in the counterclockwise direction.

The present section culminates in the next theorem. The bijection established there parametrizes the measures $\gamma \in \Gamma_G(\rho_+, \rho_-)$ without crossings using the feasible potentials ϕ on G. A preliminary lemma paves the way. The boundary conditions for feasibility merely ensure that the flow $s := \operatorname{curl}\phi$ satisfies $\operatorname{div} s = \rho$ while making a choice for the additive constant.

Lemma 3.10. *Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and an uncrossed network G on its intervals of supply and demand. Let $X, Y : [0,t] \longrightarrow \mathbb{R}$ be the non-decreasing maps giving $\rho_+ = X_\# \lambda_{[0,t]}$ and $\rho_- = Y_\# \lambda_{[0,t]}$, extended periodically with period $t = \rho_+[\mathbb{R}]$.*

Suppose $F \in \mathrm{face}(G)$ has a vertex at $I \in \mathrm{node}(G)$ and set $p = \rho_+[(-\infty, I)]$, $n = \rho_-[(-\infty, I]]$, $r = p - n$, $a = \min\{r, r + \rho[I]\}$ and $b = \max\{r, r + \rho[I]\}$. Then feasibility of ϕ implies $\phi(F) \in [a,b]$. Moreover, $\theta \in (a,b)$ implies $X(n+\theta) \in I$ or $Y(p-\theta) \in I$, depending on whether I is an interval of supply or of demand.

Proof. Assume that ϕ is feasible, and recall that G may be visualized as a convex polygon subdivided into quadrilaterals. The face F must belong to a sequence of adjacent faces sharing the corner I, beginning and ending with the boundary faces

$A, B \in \mathrm{face}(G)$ to either side of the given node I. For example, if the given node were I_0 in figure 4*a*, the sequence of faces would be $A = F_5, F_3, F_2, F_1, F_6 = B$. The arcs between these faces all point in the same direction relative to I: toward or away, depending on the sign of $\rho[I]$. The condition $\mathrm{curl}\,\phi \geqslant 0$ implies the values of ϕ vary monotonically along this sequence, so taking $\phi(A) \leqslant \phi(B)$ yields $\phi(F) \in [\phi(A), \phi(B)]$. Noting (3.3), feasibility of ϕ prescribes these boundary values to be $\phi(A) = a$ and $\phi(B) = b$.

Furthermore, suppose I is an interval of supply so that $a = p - n$ while $b = p - n + \rho_+[I]$. Then $a < \theta < b$ means that $n + \theta$ lies strictly between $p = \rho_+[(\infty, I)]$ and $p + \rho_+[I]$, which yields $X(n+\theta) \in I$. Alternatively, if I is an interval of demand, then $a = p - n - \rho_-[I]$ while $b = p - n$. In this case $-b < -\theta < -a$ forces $p - \theta$ strictly between $n = \rho_-[(-\infty, I)]$ and $n + \rho_-[I]$, which yields $Y(p - \theta) \in I$. ■

Theorem 3.11. *Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and an uncrossed network G on its intervals of supply and demand. The measures $\gamma \in \Gamma(\rho_+, \rho_-)$ compatible with G correspond bijectively to the feasible potentials ϕ on G. This bijection is encoded by the formula*

$$\mathrm{curl}\,\phi(I, J) = \gamma[I \times J] \quad \textit{on } \mathrm{arc}(G). \tag{3.4}$$

More explicitly, the restriction $\gamma|_{I\times J}$ has $X_\#\lambda_{[n+\phi(L),n+\phi(R)]}$ and $Y_\#\lambda_{[p-\phi(R),p-\phi(L)]}$ for marginals, where $X, Y : [0, t] \longrightarrow \mathbb{R}$ are the non-decreasing maps giving $\rho_+ = X_\#\lambda$ and $\rho_- = Y_\#\lambda$†. Here $n = \rho_-[(-\infty, I)]$, $p = \rho_+[(-\infty, J)]$ and $t = \rho_+[\mathbb{R}]$, while $L, R \in \mathrm{face}(G)$ represent the faces to the left and right of $(I, J) \in \mathrm{arc}(G)$.

Proof. Given $\gamma \in \Gamma_G(\rho_+, \rho_-)$, proposition 3.7 shows that $s(I, J) := \gamma[I \times J]$ defines a flow $s \geqslant 0$ on $\mathrm{arc}(G)$ satisfying $\mathrm{div}\, s = \rho$. Corollary 3.5 provides a potential $\phi : \mathrm{face}(G) \longrightarrow \mathbb{R}$ with $\mathrm{curl}\,\phi = s$ to verify (3.4). Adding a suitable constant ensures feasibility of ϕ. This correspondence between γ and ϕ is one-to-one: if $\tilde{\gamma} \in \Gamma_G(\rho_+, \rho_-)$ also satisfies $\mathrm{curl}\,\phi(I, J) = \tilde{\gamma}[I \times J]$ on $\mathrm{arc}(G)$, then proposition 3.7 combines with (3.4) to yield $\tilde{\gamma} = \gamma$.

Conversely, given a feasible potential ϕ it remains to construct $\gamma \in \Gamma_G(\rho_+, \rho_-)$ that verifies (3.4). We define

$$\gamma := \sum_{(I,J)\in \mathrm{arc}(G)} \gamma|_{I\times J}, \tag{3.5}$$

where $\gamma|_{I\times J}$ is the measure constructed in proposition 2.12 to have no crossings and with marginals $X_\#\lambda_{[n+\phi(L),n+\phi(R)]}$ and $Y_\#\lambda_{[p-\phi(R),p-\phi(L)]}$. We first need to verify that $X_\#\lambda_{[n+\phi(L),n+\phi(R)]}$ and $Y_\#\lambda_{[p-\phi(R),p-\phi(L)]}$ assign their full masses to the disjoint intervals I and J. This not only ensures that $\gamma|_{I\times J}$ is well defined, but also that it coincides with the restriction of (3.5) to $I \times J$. (Recall that the nodes of G decompose $\mathring{\mathbb{R}}$ into disjoint intervals.) The mass of this restriction is given by its marginals to be $\phi(R) - \phi(L) = \mathrm{curl}\,\phi(I, J)$, so γ verifies (3.4).

To see that $X_\#\lambda_{[n+\phi(L),n+\phi(R)]}$ assigns full mass to I, observe $\phi(L) \leqslant \phi(R)$ follows from the feasibility condition $\mathrm{curl}\,\phi \geqslant 0$. Since the interval I of supply forms a corner of L and of R, lemma 3.10 yields $X(n+\theta) \in I$ whenever $\phi(L) < \theta < \phi(R)$. This proves that $X_\#\lambda_{[n+\phi(L),n+\phi(R)]}$ assigns its full mass to I. The same lemma applies equally well to the interval J, which must therefore carry the full mass of $Y_\#\lambda_{[p-\phi(R),p-\phi(L)]}$.

† Extended to $\theta \in \mathbb{R}$ periodically: $X(\theta + t) = X(\theta)$ and $Y(\theta + t) = Y(\theta)$.

Before proceeding to discuss compatibility of γ with G, we also verify that the marginals μ and ν of γ satisfy $\mu = \rho_+$ and $\nu = \rho_-$. Since $I \in \mathrm{node}(G)$ is an interval of supply, it is connected by arcs $(I, J_i) \in \mathrm{arc}(G)$ to $j \leqslant m+1$ intervals of demand $J_1, \ldots, J_j$, enumerated clockwise around the circle starting and ending adjacent to I. The two adjacent intervals of demand share arcs with I because of maximality of G: definition 3.1(iv). If the faces to the left and right of (I, J_i) are denoted by $L_i, R_i \in \mathrm{face}(G)$, then the interior faces, $R_i = L_{i+1}$, coincide for $i = 1, 2, \ldots, j-1$, while L_1 is the boundary face between I and J_1 and R_j is the boundary face between I and J_j. Feasibility ensures that the value of ϕ is non-decreasing as one passes from each face to the next one with a corner at I, starting with L_1 and ending with R_j. Now the restriction of μ to I must be given by the sum of $X_{\#}\lambda_{[n+\phi(L_i),n+\phi(R_i)]}$ over $i = 1, \ldots, j$. Since

$$\phi(R_i) = \phi(L_{i+1}) \leqslant \phi(R_{i+1}),$$

this yields

$$\mu|_I = X_{\#}\lambda_{[n+\phi(L_1),n+\phi(R_j)]}.$$

The boundary values

$$n + \phi(L_1) = \rho_+[(-\infty, I)] \quad \text{and} \quad \phi(R_j) - \phi(L_1) = \rho[I],$$

give $\mu|_I = \rho|_I$ as desired. Now I was an arbitrary interval of supply, so $\mu = \rho_+$ is satisfied; $\nu = \rho_-$ can be proved in the same way.

Compatibility of γ with G amounts to verifying definition 3.2. Since $\gamma[I \times J] = 0$ for $(I, J) \in \mathrm{node}(G)^2 \setminus \mathrm{arc}(G)$ is manifest in (3.5), only the no-crossing property of $\mathrm{spt}\,\gamma$ need be addressed. This property respects the operation of set closure, so it suffices to show that $O(x, y)$ and $O(x', y')$ do not intersect non-tangentially for any pair of points in the set $\Omega := \cup_{\mathrm{arc}(G)}(I \times J) \cap \mathrm{spt}\,\gamma|_{I\times J}$ of full mass. Therefore, choose (x, y) and (x', y') from Ω, say $(x, y) \in I \times J$ and $(x', y') \in I' \times J'$ with (I, J) and (I', J') in $\mathrm{arc}(G)$.

To derive a contradiction, suppose $O(x, y)$ and $O(x', y')$ intersect non-tangentially. Then the points x, x', y, y' must be distinct and ordered either clockwise or counterclockwise around the circle; assume clockwise without loss. Clearly $(I, J) \neq (I', J')$, for otherwise both (x, y) and (x', y') would belong to the set $\mathrm{spt}\,\gamma|_{I\times J}$, which, by construction, has no crossings. If all four intervals I, I', J, J' were distinct, then they too would be ordered clockwise on the circle; but since (I, J) and (I', J') are arcs of G this contradicts the assumption that the network G be uncrossed. The only remaining possibilities are $I = I'$ but $J \neq J'$ (and the case $I \neq I'$ but $J = J'$, which can be handled similarly). Let L' and $R' \in \mathrm{face}(G)$ denote the faces to the left and right of (I, J'). Since I, J, J' are ordered clockwise on the circle, the considerations above show that $\phi(R) \leqslant \phi(L')$. Thus one proceeds counterclockwise in I from $x \in \mathrm{spt}\, X_{\#}\lambda_{[n+\phi(L),n+\phi(R)]}$ through $X(n + \phi(R))$ to $x' \in \mathrm{spt}\, X_{\#}\lambda_{[n+\phi(L'),n+\phi(R')]}$. This contradicts the clockwise ordering supposed for x, x', y, y', thereby establishing the theorem. ■

Corollary 3.12. *Take ρ, G, γ and ϕ from theorem 3.11 and $c(x, y)$ Borel on $\mathbb{R}^2$. Adopting the same notation as in the theorem,* $\mathrm{curl}\,\phi = \gamma$ *implies*

$$\int_{I\times J} c\,\mathrm{d}\gamma = \int_{\phi(L)}^{\phi(R)} c(X(n+\theta), Y(p-\theta))\,\mathrm{d}\theta. \tag{3.6}$$

Proof. The restriction of γ to $I \times J$ has marginals $\mu := X_{\#}\lambda_{[n+\phi(L),n+\phi(R)]}$ and $\nu := Y_{\#}\lambda_{[p-\phi(R),p-\phi(L)]}$ by theorem 3.11. When $m > 0$, it is clear that $X(n+\theta)$ moves counterclockwise around the circle as θ is increased, sweeping through I from one end to the other as θ moves from $\phi(L)$ to $\phi(R)$. At the same time, $Y(p-\theta)$ sweeps clockwise through J. (When $m = 0$, these same facts follow directly from feasibility, which implies $\phi(L) = \rho_+[(-\infty, I)] - n$ while $\phi(R) = p - \rho_-[(-\infty, J)]$.) In both cases, remark 2.13 shows that the curve $Z(\theta) := (X(n+\theta), Y(p-\theta))$ defined on $\phi(L) < \theta < \phi(R)$ has no crossings. Neither will the measure $Z_{\#}\lambda_{[\phi(L),\phi(R)]}$ in $\Gamma(\mu,\nu)$.

Now $\gamma|_{I\times J}$ has no crossings by compatibility of γ with G, and its marginals μ and ν are supported on disjoint intervals. The uniqueness assertion of proposition 2.11 forces $Z_{\#}\lambda_{[\phi(L),\phi(R)]} = \gamma|_{I\times J}$. Finally, (3.6) follows from the change of variables formula (2.7): if either integral makes sense, they both exist and coincide. ∎

4. Convex separable flow optimizations

In the preceding section, a parametrization was developed for measures γ without crossings in terms of flows (or equivalently potentials $\phi : \text{face}(G) \longrightarrow \mathbb{R}$) on uncrossed networks G. Here we take up the theme of its final corollary: exploring the properties of the transport cost associated with γ as a function of the values of ϕ.

Given an uncrossed network G on the supply and demand intervals of $\rho \in \mathcal{M}_0^m(\mathbb{R})$, our first lemma shows that the feasible potentials ϕ form a compact convex polytope. Theorem 4.4 goes on to assert that the transport cost $\mathcal{C}(\gamma)$ associated with ϕ must be convex and separable (1.4) as a function of the variables $\phi(F)$ with $F \in \text{face}(G)$. Thus, for costs of concave type, the transportation problem reduces to the optimization of several convex separable network flows: the infinite-dimensional linear problem is replaced by ($3m$ choose m)/($2m+1$) convex minimizations, each in m dimensions. Because of separability, excellent algorithms exist for computing the optimal flows (see Rockafellar 1984). Small examples can be solved explicitly.

Lemma 4.1. *Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and an uncrossed network G on its intervals of supply and demand. Then the feasible potentials ϕ form a compact convex set: enumerating interior faces $F_1, \ldots, F_m \in \text{face}(G)$, the map $(\phi_1, \ldots, \phi_m) := (\phi(F_1), \ldots, \phi(F_m))$ defines a bijection from the feasible potentials onto a convex polytope $\Phi_G \subset \mathbb{R}^m$.*

Proof. Define $\Phi_G := \{(\phi(F_1), \ldots, \phi(F_m)) \in \mathbb{R}^m \mid \phi \text{ is a feasible potential on } G\}$. The feasible potentials $\phi : \text{face}(G) \longrightarrow \mathbb{R}$ correspond bijectively with the points of Φ_G because their values on boundary faces of G are prescribed by ρ in definition 3.8. The only other requirement for feasibility is non-negativity of the flow curl ϕ, meaning $\phi(L) \leqslant \phi(R)$ holds whenever the faces L and R lie adjacent to the left and the right of an arc in G. Thus, Φ_G is defined by finitely many linear inequalities—one for each of the $3m+1$ arcs in G—making it a convex polytope. Since Φ_G is closed, compactness is implied by lemma 3.10, which asserts that the maximum and minimum values of ϕ are attained on boundary faces of G. ∎

Now fix an interior face $F \in \text{face}(G)$ and consider how the transport cost depends on $\phi(F)$. Since F is bounded by four arcs of G, this value of the potential occurs in four terms (3.6) contributing to $\mathcal{C}(\gamma)$; two increase with $\phi(F)$ while two decrease. Remarkably, the assumption that $c(x,y)$ be of concave type precisely ensures that

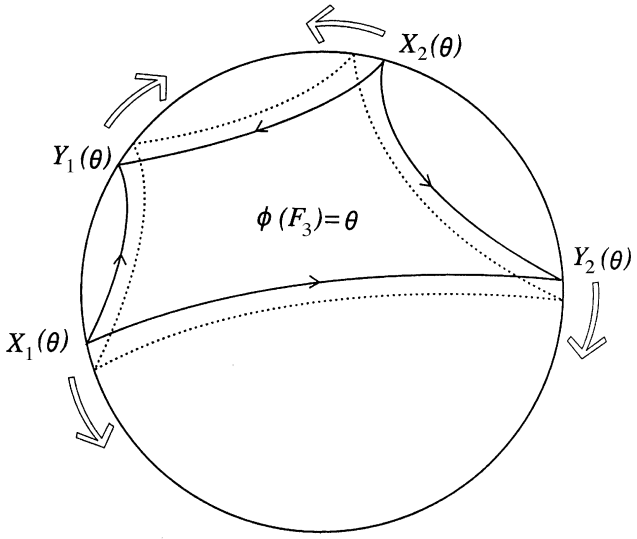

Figure 7. Dependence of the transport cost on $\phi(F_i)$ is probed by varying θ.

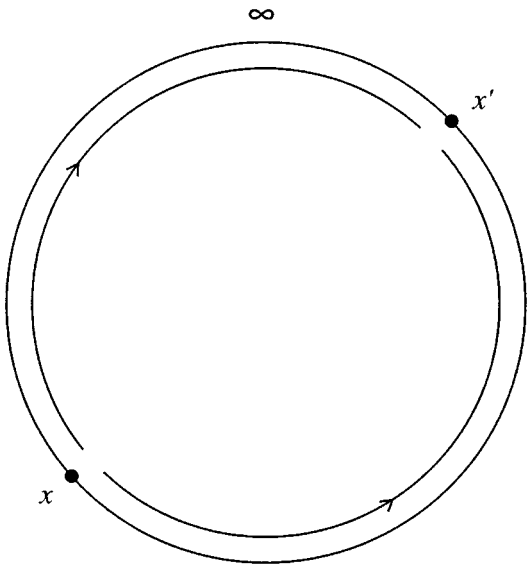

Figure 8. Concave type: means $c(x,\cdot) - c(x',\cdot)$ increase as $\cdot$ slides from x toward x'.

they combine to make the transport cost a convex function of $\phi(F)$. This is proved in theorem 4.4, essentially by differentiating the transport cost with respect to $\phi(F)$. A proposition first establishes monotonicity of the putative derivative, which, as figure 7 suggests, is given by (4.1). Lemma B 1 (figure 8) provides a key ingredient.

Proposition 4.2. *Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$, an uncrossed network, G, on its intervals of supply and demand and a cost of concave type, $c(x,y)$. If $F \in \mathrm{face}(G)$ is an interior face, two of its four boundary arcs run clockwise: call them (I_1, J_1) and $(I_2, J_2) \in$*

arc(G). *A non-decreasing function,* $f:(a,b)\longrightarrow\mathbb{R}$, *is defined by*

$$f(\theta)=c(X(n_1+\theta),Y(p_1-\theta))+c(X(n_2+\theta),Y(p_2-\theta))\\ -c(X(n_1+\theta),Y(p_2-\theta))-c(X(n_2+\theta),Y(p_1-\theta)); \tag{4.1}$$

here

$$a=\max_j\{\rho[(-\infty,I_j)],\rho[(-\infty,J_j]]\},\qquad b=\min_j\{\rho[(-\infty,I_j]],\rho[(-\infty,J_j)]\},$$
$$n_j:=\rho_-[(-\infty,I_j)]\quad\textit{and}\quad p_j:=\rho_+[(-\infty,J_j)],\quad\textit{for } j=1,2,$$

while X *and* $Y:[0,t]\longrightarrow\mathbb{R}$ *are the non-decreasing maps with* $\rho_+=X_\#\lambda$ *and* $\rho_-=Y_\#\lambda$, *extended periodically* $X(\theta)=X(\theta+t)$ *and* $Y(\theta)=Y(\theta+t)$ *with period* $t=\rho_+[\mathbb{R}]$.

Moreover, f *increases strictly unless* ρ_+ *has point masses at both* $X(n_j+\theta)$, $j=1,2$, *while* ρ_- *has point masses at both locations* $Y(p_j-\theta)$ *for the same* $\theta\in(a,b)$.

Proof. Assume $a<b$, since otherwise the proposition is vacuous. Before going further, we observe from lemma 3.10 that $a<\theta<b$ implies $X(n_j+\theta)\in I_j$ and $Y(p_j-\theta)\in J_j$ for $j=1,2$.

Now consider the cross-difference

$$\Delta(x_1,y_1,x_2,y_2):=c(x_1,y_1)+c(x_2,y_2)-c(x_1,y_2)-c(x_2,y_1), \tag{4.2}$$

defined on $I_1\times J_1\times I_2\times J_2$. Since the arcs (I_j,J_j) run clockwise around the boundary of F, the four disjoint intervals I_1, J_1, I_2 and J_2 are ordered clockwise on the circle $\mathring{\mathbb{R}}\approx\boldsymbol{S}^1$. Remark 4.3 notwithstanding, (4.2) takes only real values for costs of concave type since the remark before lemma B 1 (in Appendix B) shows $c(x,y)>-\infty$ if $x\neq y$. Fixing any three variables (say x_1, y_1 and x_2), corollary B 3 shows (4.2) to be strictly monotone in the fourth variable (in this case $y_2\in J_2$). More specifically, the ordering implies that $\Delta(x_1,y_1,x_2,y_2)$ is increased when either y_1 or y_2 is moved clockwise, or when x_1 or x_2 moves counterclockwise. Monotonicity of X and Y makes it clear that increasing θ moves $y_j=Y(p_j-\theta)$ clockwise through J_j and $x_j=X(n_j+\theta)\in I_j$ counterclockwise, as indicated in figure 7. Now fix $a<\theta<\theta'<b$ and set $y_j'=Y(p_j-\theta')$ and $x_j'=X(n_j+\theta')$. The difference

$$\begin{aligned}f(\theta')-f(\theta)&=\Delta(x_1',y_1',x_2',y_2')-\Delta(x_1,y_1',x_2',y_2')\\&\quad+\Delta(x_1,y_1',x_2',y_2')-\Delta(x_1,y_1,x_2',y_2')\\&\quad+\Delta(x_1,y_1,x_2',y_2')-\Delta(x_1,y_1,x_2,y_2')\\&\quad+\Delta(x_1,y_1,x_2,y_2')-\Delta(x_1,y_1,x_2,y_2)\end{aligned}$$

is expressed as a sum of four non-negative quantities, verifying f non-decreasing on (a,b). Moreover, $f(\theta')>f(\theta)$ unless all four terms vanish, in which case $X(n_j+\theta')=X(n_j+\theta)$ and $Y(p_j-\theta')=Y(p_j-\theta)$. Thus f increases strictly, unless mass $\theta'-\theta$ is concentrated by ρ_+ at both x_1 and x_2, and by ρ_- at both y_1 and y_2. ■

Remark 4.3. Technically, it is possible that $X(\theta)$ or $Y(\theta)$ is infinite at $\theta=0$ and t. Should ∞ also lie in the interval I_j or J_j, $j=1,2$, then (4.1) may fail to be well defined for one value of θ in (a,b). The convex function (4.3), however, remains well defined.

To each interior face $F_i \in \mathrm{face}(G)$, proposition 4.2 associates a non-decreasing function $f_i : (a_i, b_i) \longrightarrow \mathbb{R}$ through (4.1). A convex function $C_i(s)$ is then defined on $[a_i, b_i]$ by

$$C_i(s) = \int_{z_i}^{s} f_i(\theta)\,\mathrm{d}\theta. \tag{4.3}$$

Here $z_i := \frac{1}{2}(a_i + b_i)$ is chosen to be the centre of $[a_i, b_i]$, so that $C_i(s)$ takes finite real values except perhaps at $s = a_i$ and $s = b_i$.

Since we also want the transport cost $\mathcal{C}(\gamma)$ to take finite values on $\Gamma(\rho_+, \rho_-)$, it is convenient to assume the excess production ρ satisfies the 'moment conditions'

$$\int c(x, q)\,\mathrm{d}\rho_+(x) < \infty \quad \text{and} \quad \int c(q, y)\,\mathrm{d}\rho_-(y) < \infty, \tag{4.4}$$

for some (and hence all) $q \in \mathbb{R}$.

Theorem 4.4. *Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$, an uncrossed network G on its intervals of supply and demand, and a Borel cost of concave type $c(x, y) \geqslant 0$ satisfying (4.4). To each feasible potential ϕ associate the transport cost $\mathcal{C}_G(\phi) := \mathcal{C}(\gamma)$, where $\operatorname{curl}\phi = \gamma$ for $\gamma \in \Gamma_G(\rho_+, \rho_-)$. Enumerate the interior faces $F_1, \ldots, F_m \in \mathrm{face}(G)$. Then for some constant $C_0 \in \mathbb{R}$,*

$$\mathcal{C}_G(\phi) = C_0 + \sum_{i=1}^{m} C_i(\phi(F_i)), \tag{4.5}$$

where $C_i(s)$ denotes the convex function (4.3) corresponding to the face F_i.

Proof. Define $p(I) := \rho_+[(-\infty, I)]$ and $n(I) := \rho_-[(-\infty, I)]$ to be the total positive and negative mass (respectively) of ρ that lies between $-\infty$ and $I \in \mathrm{node}(G)$. Let $X, Y : [0, t] \longrightarrow \mathbb{R}$ denote the non-decreasing maps pushing the Lebesgue measure forward to $\rho_+ = X_{\#}\lambda_{[0,t]}$ and $\rho_- = Y_{\#}\lambda_{[0,t]}$, extended periodically to the line with period $t = \rho_+[\mathbb{R}]$. Then use

$$\epsilon_i(I, J) := \begin{cases} +1, & \text{if } (I, J) \text{ runs clockwise along the boundary of the face } F_i, \\ 0, & \text{if } (I, J) \text{ is not a boundary arc of the face } F_i, \\ -1, & \text{if } (I, J) \text{ runs counterclockwise along the boundary of } F_i, \end{cases} \tag{4.6}$$

to define the function

$$K(\phi_1, \ldots, \phi_m) := \sum_{i=1}^{m} \sum_{(I,J)\in \mathrm{arc}(G)} \epsilon_i(I, J) \int_0^{\phi_i} c(X(n(I) + \theta), Y(p(J) - \theta))\,\mathrm{d}\theta, \tag{4.7}$$

on the rectangle $\Pi := [a_1, b_1] \times \cdots \times [a_m, b_m]$. Here a_i and b_i are chosen so that $[a_i, b_i] = \bigcap \mathrm{conv}\{p(I) - n(I), p(I) - n(I) + \rho[I]\}$, the intersection being over the four corners $I \in \mathrm{node}(G)$ of the quadrilateral F_i; $\mathrm{conv}\,K$ denotes the convex hull of $K \subset \mathbb{R}$.

Our first observation is that the rectangle Π contains the feasible set $\Phi_G \subset \mathbb{R}^m$; i.e. if $\phi : \mathrm{face}(G) \longrightarrow \mathbb{R}$ is feasible, then $\phi(F_i) \in [a_i, b_i]$ for $i = 1, \ldots, m$. This follows directly from the definition of $[a_i, b_i]$ and lemma 3.10.

Our next step is to verify that the transport cost $\mathcal{C}_G(\phi)$ coincides with the restriction of $\mathcal{K}$ to Φ_G: given a feasible potential $\phi : \mathrm{face}(G) \longrightarrow \mathbb{R}$, or the measure

$\gamma \in \Gamma_G(\rho_+, \rho_-)$ with $\operatorname{curl}\phi = \gamma$ from theorem 3.11, the claim to be established is that $\mathcal{C}(\gamma) = \mathcal{K}(\phi(F_1), \dots, \phi(F_m))$. From the moment conditions (4.4) and lemma B 6, the integrals defining $\mathcal{K}$ converge and the order of the sums can be interchanged in (4.7). Associate to $(I, J) \in \operatorname{arc}(G)$ the indices $R(I, J)$ and $L(I, J)$ of the faces $F_{R(I,J)}$ and $F_{L(I,J)}$ that lie immediately to the right and left of the arc (I, J). The observation that $\epsilon_i(I, J)$ is positive for $i = R(I, J)$ and negative for $i = L(I, J)$, vanishing otherwise, allows us to compute the sum over $i = 1, \dots, m$ in (4.7):

$$\mathcal{K}(\phi_1, \dots, \phi_m) = \sum_{(I,J)\in\operatorname{arc}(G)} \int_{\phi_{L(I,J)}}^{\phi_{R(I,J)}} c(X(n(I)+\theta), Y(p(J)-\theta))\,\mathrm{d}\theta. \tag{4.8}$$

Recalling that $\gamma \in \Gamma_G(\mu, \nu)$ vanishes outside of the disjoint union $\bigcup_{(I,J)\in\operatorname{arc}(G)} I \times J$, comparing (4.8) with (3.6) yields $\mathcal{K}(\phi(F_1), \dots, \phi(F_m)) = \mathcal{C}(\gamma)$.

Having shown that $\mathcal{C}_G(\phi) = \mathcal{K}(\phi(F_1), \dots, \phi(F_m))$, what remains to be verified is that $\mathcal{K}(\phi_1, \dots, \phi_m) - \sum_{i=1}^m C_i(\phi_i)$ takes a constant value throughout the rectangle Π. Interchanging the order of the sum and the integral in (4.7) yields

$$\mathcal{K}(\phi_1, \dots, \phi_m) := \sum_{i=1}^m \int_0^{\phi_i} f_i(\theta)\,\mathrm{d}\theta, \tag{4.9}$$

where the integrands $f_i(\theta)$ are defined for $\theta \in \mathbb{R}$ by

$$f_i(\theta) := \sum_{(I,J)\in\operatorname{arc}(G)} \epsilon_i(I, J) c(X(n(I)+\theta), Y(p(J)-\theta)). \tag{4.10}$$

Fix $i \in \{1, \dots, m\}$, and let I_1, J_1, I_2, J_2 denote the four nodes at the corners of the face F_i, labelled clockwise around the circle starting with an interval of supply. Evaluating $f_i(\theta)$ from (4.10) and the definition (4.6) of $\epsilon_i(I, J)$ yields

$$\begin{aligned} f_i(\theta) = {} & c(X(n(I_1)+\theta), Y(p(J_1)-\theta)) + c(X(n(I_2)+\theta), Y(p(J_2)-\theta)) \\ & - c(X(n(I_1)+\theta), Y(p(J_2)-\theta)) - c(X(n(I_2)+\theta), Y(p(J_1)-\theta)). \end{aligned} \tag{4.11}$$

Comparison with (4.1) makes it clear that the functions $f_i(\theta)$ appearing in (4.9) and (4.3) are the same, at least on $[a_i, b_i]$ where both are defined. Thus

$$\mathcal{K}(\phi_1, \dots, \phi_m) = C_0 + \sum_{i=1}^m C_i(\phi_i),$$

holds with the constant

$$C_0 := \sum_{i=1}^m \int_0^{z_i} f_i(\theta)\,\mathrm{d}\theta.$$

These integrals converge by lemma B 6 again. ■

To summarize, the last two sections imply that minimizing the transport cost $\mathcal{C}(\gamma)$ among measures $\gamma \in \Gamma(\rho_+, \rho_-)$ compatible with G is equivalent to minimizing the convex function

$$\mathcal{C}_G(\phi_1, \dots, \phi_m) := \sum_{i=1}^m C_i(\phi_i), \tag{4.12}$$

on a convex polytope $\Phi_G \subset \mathbb{R}^m$. The left and right partial derivatives of $\mathcal{C}_G$ may be computed from (4.3) to be

$$\frac{\partial \mathcal{C}_G}{\partial \phi_i^{\pm}}(\phi_1, \ldots, \phi_m) = \lim_{\theta \to \phi_i^{\pm}} f_i(\theta), \tag{4.13}$$

where the non-decreasing function $f_i(\phi_i)$ associated with the face F_i is given by (4.1). Geometrically, it is the spatial variables rather than the mass variables ϕ_i that play the central role: they represent the four points $x_j = X(n_j + \phi_i)$ and $y_j = Y(p_j - \phi_i)$, $j = 1, 2$, at which the measure $\gamma = \operatorname{curl} \phi$ divides the flow of mass from the interval of supply I_j into the intervals of demand J_1 and J_2 (compare theorem 3.11 with figure 7). The stationarity condition for the cost pits the cross-difference $f_i(\phi_i)$ of the four points x_1, y_1, x_2 and y_2 against the feasibility constraints. At last, this puts us in a position to verify example 1.1 as follows.

Example 1.1 (revisited) Let ρ be distributed sinusoidally over -10 to 10: $\mathrm{d}\rho(x) = \sin(\pi x/5)\,\mathrm{d}x$. Then $m = 1$, yielding a unique uncrossed network G: a square with vertices in the intervals $I_1 := [-10, -5] = -J_1$ and $I_2 = [0, 5] = -J_2$. The boundary values of ϕ for feasibility are $0, (10/\pi), 0, (10/\pi)$, while $\phi(F_1)$ must lie in $\Phi_G = [0, (10/\pi)]$. To minimize the convex cost $\mathcal{C}_G(\phi_1)$ on Φ_G, use the symmetries

$$X(\theta) + 10 = X(n_2 + \theta) = -Y(p_1 - \theta) = 10 - Y(p_2 - \theta),$$

to compute its derivative from (4.1) and (4.13)

$$\frac{\mathrm{d}\mathcal{C}_G}{\mathrm{d}\phi_1} = c(x - 10, 10 - x) + c(x, -x) - c(x - 10, -x) - c(x, 10 - x), \tag{4.14}$$

in terms of $x = X((10/\pi) + \phi_1)$. For the concave cost $c(x, y) := \sqrt{2|x - y|}$, the minimum of $\mathcal{C}_G$ occurs at $x = 1$ when (4.14) vanishes. Since the optimal measure $\gamma \in \Gamma(\rho_+, \rho_-)$ must be compatible with G (theorem 2.5 and lemma 3.3), this confirms the description (1.2)–(1.3) of the optimal map $s(x)$ and measure γ.

5. Global optimization and separation of scales

This fifth and final section is devoted to a characterization of optimal measures that shows the sufficiency of three necessary conditions as follows.

Theorem 5.1. *Fix $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and a continuous cost $c(x, y) \geqslant 0$ of concave type. A measure $\gamma \in \Gamma(\rho_+, \rho_-)$ is optimal provided:*

(i) *γ has no crossings;*

(ii) *$\mathcal{C}(\gamma) < \infty$; and*

(iii) *$\mathcal{C}(\gamma) \leqslant \mathcal{C}(\tilde{\gamma})$ whenever γ and $\tilde{\gamma} \in \Gamma(\rho_+, \rho_-)$ are both compatible with the same uncrossed network G on ρ's intervals of supply and demand.*

The beauty of this result is that it combines with the preceding sections to yield a powerful algorithm for computing and verifying exact solutions of the transportation problem. Indeed, optimality of an uncrossed measure γ can now be verified with a finite computation. The present theorem asserts that γ is optimal if and only if its

cost is a minimum among all $\tilde{\gamma} \in \Gamma_G(\rho_+, \rho_-)$ and uncrossed networks G compatible with γ. Parametrizing $\Gamma_G(\rho_+, \rho_-)$ in the coordinates of the preceding section, the transport cost is given by a (separable) convex function $\mathcal{C}_G(\phi_1, \ldots, \phi_m)$ on a polytope $\Phi_G \subset \mathbb{R}^m$. Convexity ensures that any critical point will be a minimum of $\mathcal{C}_G$ on Φ_G. Thus the conditions guaranteeing optimality are local in ϕ: γ is optimal if and only if the corresponding point $(\phi_1, \ldots, \phi_m)$ satisfies the Kuhn–Tucker equations for a critical point on each uncrossed network G compatible with γ. There are only finitely many such networks to check, while the explicit form of the equations may be recovered from (4.1), (4.13) and the constraints of definition 3.8.

Similarly, to find an optimal measure, choose any $\gamma \in \Gamma(\mu, \nu)$ having no crossings. Unless γ is optimal, it is compatible with some uncrossed network G on which the transport cost can be lowered. Minimize the restriction $\mathcal{C}_G(\phi_1, \ldots, \phi_m)$ of this cost to $\Phi_G \approx \Gamma_G(\rho_+, \rho_-)$ using one of the network flow algorithms referred to in § 4. Replace γ by the minimizing measure $\tilde{\gamma} \approx d^*\phi$, and repeat this procedure to obtain an optimal measure in no more than d_m iterations. As with the simplex algorithm, one may hope that the required iterations number far fewer than d_m, since one need not check networks G with which γ is incompatible.

The idea underlying theorem 5.1 is that optimality of γ can be established by testing independently on m different size scales, as long as each pair of neighbouring scales is consistent. Viability of this approach is hinted at by the hierarchical structure of uncrossed networks as well as by the rule of three. Before delving into the argument, we recall (without proof) a characterization of optimal measures based on inequalities related to (2.1). First derived by Smith & Knott (1992) from duality-based work of Rüschendorf (1991), a direct demonstration of necessity is given in Gangbo & McCann (1996, thm 2.3); sufficiency can then be inferred from uniqueness.

Definition 5.2. A subset $\Omega \subset \mathbb{R}^2$ is called *c-cyclically monotone* if, for every finite collection of points $(x_i, y_i) \in \Omega$, $i = 1, \ldots, n$, setting $y_{n+1} = y_1$ yields

$$\sum_{i=1}^{n} c(x_i, y_i) \leqslant \sum_{i=1}^{n} c(x_i, y_{i+1}). \tag{5.1}$$

Theorem 5.3 (Smith & Knott). *Let $c(x, y) \geqslant 0$ be continuous, $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ and $\mathcal{C}(\gamma) < \infty$. Then γ is optimal if and only if* $\operatorname{spt}\gamma$ *is c-cyclically monotone.*

As the next proposition shows, c-cyclical monotonicity can be used to deduce a consistency condition matching optimal solutions on different domains. Before proving the proposition, we state a lemma that demonstrates that any set without crossings is contained in complementary squares on the torus $\mathring{\mathbb{R}} \times \mathring{\mathbb{R}} \sim \boldsymbol{S}^1 \times \boldsymbol{S}^1$.

Lemma 5.4. *Suppose $T \subset \mathbb{R}^2$ has no crossings. Choose $(a, b) \in T$, and let $I \subset \mathbb{R}$ denote the compact interval with end-points a and b; denote the closure of its complement by $\tilde{I} :=$* closure$(\mathbb{R} \setminus I)$. *Then $T \subset (I \times I) \cup (\tilde{I} \times \tilde{I})$.*

Proof. Let $(x, y) \in T$. Then either $x \in I$ or $x \in \tilde{I}$. If both are true, i.e. $x = a$ or b, then certainly $(x, y) \in (I \times I) \cup (\tilde{I} \times \tilde{I})$. On the other hand, since the circle $O(x, y)$ does not cross a circle $O(a, b)$ through the end-points of I, an interior point $x \in I$ forces $y \in I$; similarly $x \notin I$ forces $y \in \tilde{I}$. Either way, $(x, y) \in I \times I$ or $\tilde{I} \times \tilde{I}$. ∎

Proposition 5.5. *Choose a cost $c(x,y)$ of concave type and a compact interval $I \subset \mathbb{R}$ with end-points a and b. Let $S \subset \mathbb{R}^2$ be contained in the square $I \times I$, while $T \subset \mathbb{R}^2$ is disjoint from this square. Then c-cyclical monotonicity of $S \cup \{(a,b)\}$ and $T \cup \{(a,b)\}$ implies the same for their union $S \cup T$.*

Proof. Assume $(a,b) \in S$ without loss of generality. Also assume $a \leqslant b$; the other case is handled by observing $\tilde{c}(x,y) := c(-x,-y)$ to be of concave type, while the reflections of S and T through the origin are $\tilde{c}$-cyclically monotone sets. Finally, observe c-cyclical monotonicity precludes crossings in $T \cup \{(a,b)\}$, by the two-point $n = 2$ inequality (5.1) and the definition of concave type. Thus, lemma 5.4 yields $T \subset \tilde{I} \times \tilde{I}$, where $\tilde{I}$ denotes the closure of the complement of $I \subset \mathbb{R}$.

To any finite sequence of points $(x_1,y_1),\dots,(x_n,y_n) \in S \cup T$, associate an integer $s \leqslant \frac{1}{2}n$ that counts the number of *switches*: that is, the number of $i = 1,\dots,n$ for which $(x_i,y_i) \in S$ but $(x_{\sigma(i)},y_{\sigma(i)}) \in T$; here $\sigma = (12\dots n)$ denotes the cyclic permutation on n letters. We shall establish (5.1) by induction on s.

When $s = 0$ the complete sequence of points is contained either in S or T, so (5.1) follows from c-cyclical monotonicity of S or of T. If $s > 0$ the sequence labels can be permuted cyclically to ensure $(x_n,y_n) \in S$ but $(x_1,y_1) \in T$ without affecting the veracity of (5.1). Now let $j < n$ denote the smallest index for which $(x_j,y_j) \in T$ but $(x_{j+1},y_{j+1}) \in S$. An analysis of cases is required depending on the sign of the cross-difference

$$\Delta(x_j,y_{\sigma(j)},x_n,y_{\sigma(n)}) := c(x_j,y_{\sigma(j)}) + c(x_n,y_{\sigma(n)}) - c(x_j,y_{\sigma(n)}) - c(x_n,y_{\sigma(j)}). \tag{5.2}$$

Case (A) $\Delta(x_j,y_{\sigma(j)},x_n,y_{\sigma(n)}) \geqslant 0$ implies

$$\sum_1^n c(x_i,y_{\sigma(i)}) \geqslant \sum_1^j c(x_i,y_{\tau(i)}) + \sum_{j+1}^n c(x_i,y_{\pi(i)}) \geqslant \sum_1^n c(x_i,y_i).$$

Here $\tau = (12\dots j)$ and $\pi = (j+1\ j+2\dots n)$ are cyclic permutations of two disjoint subsequences and the second inequality follows from the inductive hypothesis: the subsequence of length j has zero switches since its elements lie entirely in T, while the subsequence of length $n-j$ begins and ends in S so has $s-1$ switches by construction.

Case (B) $\Delta(x_j,y_{\sigma(j)},x_n,y_{\sigma(n)}) < 0$ precludes any intersection of the circles $O(x_j,y_{\sigma(j)})$ and $O(x_n,y_{\sigma(n)})$ since the cost $c(x,y)$ is of concave type. Recalling $S \subset I \times I$, our choices of j and n ensure that $y_{\sigma(j)}$ and x_n lie in $I = [a,b]$, while $y_{\sigma(n)}$ and x_j lie outside of the interval (a,b). Thus the counterclockwise order of points around the circle $\mathring{\mathbb{R}}$ may be taken to be either (B1) $y_{\sigma(n)}$, a, x_n, $y_{\sigma(j)}$, b, x_j, or (B2) x_j, a, $y_{\sigma(j)}$, x_n, b, $y_{\sigma(n)}$.

As long as the ordering (B1) is respected, corollary B 3 shows the cross-difference (5.2) to increase with x_n and decrease with $y_{\sigma(j)}$. Thus,

$$\Delta(x_j,b,a,y_{\sigma(n)}) \leqslant \Delta(x_j,b,x_n,y_{\sigma(n)}) \leqslant \Delta(x_j,y_{\sigma(j)},x_n,y_{\sigma(n)}). \tag{5.3}$$

Identifying $(x_0,y_0) := (a,b)$, the sequence $(x_0,y_0),\dots,(x_j,y_j)$ lies in $T \cup \{(a,b)\}$, so c-cyclical monotonicity with $\tau = (012\dots j)$ yields

$$\sum_0^j c(x_i,y_i) \leqslant \sum_0^j c(x_i,y_{\tau(i)}). \tag{5.4}$$

The sequence $(x_{j+1}, y_{j+1}), \dots, (x_n, y_n)$ has $s-1$ switches, so, as before, the inductive hypothesis yields

$$\sum_{j+1}^{n} c(x_i, y_i) \leqslant \sum_{j+1}^{n} c(x_i, y_{\pi(i)}), \tag{5.5}$$

where $\pi = (j+1\ j+2 \dots n)$. Summing (5.3)–(5.5) gives the desired conclusion, i.e. (5.1).

Similarly, if the ordering (B2) is respected, then corollary B 3 shows (5.2) to increase with $y_{\sigma(n)}$ and decrease with x_j in $\mathring{\mathbb{R}}$. Thus,

$$\Delta(a, y_{\sigma(j)}, x_n, b) \leqslant \Delta(x_j, y_{\sigma(j)}, x_n, y_{\sigma(n)}). \tag{5.6}$$

Identifying

$$(x_{n+1}, y_{n+1}) := (a, b) \in S,$$

the inductive hypothesis yields c-cyclical monotonicity of the sequence

$$(x_1, y_1), \dots, (x_j, y_j),$$

and also of

$$(x_{j+1}, y_{j+1}), \dots, (x_{n+1}, y_{n+1});$$

the first sequence has zero switches, while the second sequence has $s-1$. Replacing the lower limits in (5.4) by $i = 1$ and the upper limits in (5.5) by $i = n+1$, summing (5.4)–(5.6) with $\tau = (12 \dots j)$ and $\pi = (j+1\ j+2 \dots n+1)$ yields (5.1) to complete the proof. ■

After a preliminary lemma, one more inductive argument will complete the proof of theorem 5.1.

Lemma 5.6. *Suppose $\gamma \in \mathcal{M}_+(\mathbb{R}^2)$ has no crossings, and its mass lies on the product of two disjoint intervals $I, J \subset \mathring{\mathbb{R}}$. For some compact interval $K \subset \mathbb{R}$, with complement $\tilde{K}$ and end-points $a \in I$ and $b \in J$, $\gamma = \gamma^0 + \gamma^1$ decomposes into non-negative measures with $\operatorname{spt}\gamma^1 \subset K \times K$ and $\operatorname{spt}\gamma^0 \subset \operatorname{closure}(\tilde{K} \times \tilde{K})$, but $(a, b) \in \operatorname{spt}\gamma^0 \cap \operatorname{spt}\gamma^1$.*

Proof. For any $(a, b) \in \operatorname{spt}\gamma$, if γ^1 is defined as the restriction of γ to $K \times K$, then lemma 5.4 will guarantee $\operatorname{spt}\gamma^0 \subset \operatorname{closure}(\tilde{K} \times \tilde{K})$. However, to ensure $(a, b) \in \operatorname{spt}\gamma^0 \cap \operatorname{spt}\gamma^1$ requires a brief argument.

Corollary 2.12 represents $\gamma = Z_{\#}\lambda_{[0,t]}$ as the image of a Lebesgue measure on a curve $Z(\theta) = (X(\theta), Y(\phi - \theta))$, where X and Y are non-decreasing functions on $0 < \theta < t = \gamma[\mathbb{R}]$, and $Y(\theta) := Y(\theta + t)$ extends periodically. The monotone functions, X and Y, have only countably many discontinuities, so for the rest of the proof fix a point $\theta \in (0, t)$, at which the curve Z is continuous. Let $K \subset \mathbb{R}$ be the compact interval with end-points $a := X(\theta)$ and $b := Y(\phi - \theta)$. Clearly $a \neq b$, since $a \in I$, while $b \in J$. If B denotes a small open ball centred at (a, b), then

$$Q_1 := B \cap (K \times K) \quad \text{and} \quad Q_0 := B \cap \operatorname{closure}(\tilde{K} \times \tilde{K})$$

will be, respectively, the upper-left and lower-right quadrants of B. Since Z is monotone non-increasing in a neighbourhood of $Z(\theta) = (a, b)$, the continuity of Z at θ ensures $\gamma[Q_i] > 0$ for $i = 0, 1$. Let $m \geqslant 0$ denote the mass that γ assigns to the single

point $Q_0 \cap Q_1 = \{(a,b)\}$, and define $\gamma^1 := \gamma|_{K\times K} - m\delta_{(a,b)}/2$ and $\gamma^0 := \gamma - \gamma^1$. Obviously, $\operatorname{spt}\gamma^1 \subset K \times K$, while lemma 5.4 yields $\operatorname{spt}\gamma^0 \subset \text{closure}(\tilde{K}\times\tilde{K})$. When $m > 0$, both γ^0 and γ^1 include a Dirac mass at (a,b) so certainly $(a,b) \in \operatorname{spt}\gamma^0 \cap \operatorname{spt}\gamma^1$. Otherwise $m = 0$, in which case $\gamma^i[Q_i] = \gamma[Q_i] > 0$. Since the ball B can be taken as being arbitrarily small, one must have $(a,b) \in \operatorname{spt}\gamma^i$ for $i = 0, 1$. ■

Proof of theorem 5.1. Let $c(x,y)$, $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and $\gamma \in \Gamma(\rho_+, \rho_-)$ satisfy hypotheses (i)–(iii) of theorem 5.1. To conclude that γ is optimal, we consider two cases: either $\gamma[I \times J] > 0$ for some pair of non-adjacent intervals I of supply and J of demand, or no such pair $I, J \subset \mathring{\mathbb{R}}$ of intervals exists. The latter case is easily resolved: γ would be compatible with every uncrossed network on ρ's intervals of supply and demand because definition 3.1(iv) ensures that any such network includes all arcs between adjacent nodes. Moreover, the transport cost (1.1) is known to be minimized by some optimal measure $\tilde{\gamma} \in \Gamma(\rho_+, \rho_-)$ with $\mathcal{C}(\tilde{\gamma}) \leqslant \mathcal{C}(\gamma) < \infty$. Since $\tilde{\gamma}$ has no crossings in view of theorem 2.5, it is compatible with an uncrossed network G by lemma 3.3. Hypothesis (iii) yields $\mathcal{C}(\gamma) \leqslant \mathcal{C}(\tilde{\gamma})$ whence γ is optimal.

We therefore return to the case in which $\gamma[I \times J] > 0$ for some non-adjacent intervals I of supply and J of demand. Since I would lie adjacent to J if $m \leqslant 1$, in those cases there is nothing to prove. What remains to be shown is that optimality of γ for $m > 1$ follows by induction on m.

Therefore, assume the theorem has been established for all smaller values of $m > 1$. Since γ has no crossings, neither does its restriction $\eta := \gamma|_{I\times J}$ to $I \times J$. Applying lemma 5.6 to η yields a compact interval $K \subset \mathbb{R}$ with end-points $a \in I$ and $b \in J$ and a decomposition $\eta = \eta^0 + \eta^1$ with $(a,b) \in \operatorname{spt}\eta^0 \cap \operatorname{spt}\eta^1$, where $\operatorname{spt}\eta^1 \subset K \times K$ and $\operatorname{spt}\eta^0 \subset \text{closure}(\tilde{K}\times\tilde{K})$. In view of lemma 5.4, the decomposition of $\gamma = \gamma^0 + \gamma^1$ given by $\gamma^1 := \eta^1 + (\gamma - \eta)|_{K\times K}$ inherits the same properties. We shall use the inductive hypothesis to derive optimality of γ^0 and γ^1 with respect to the cost function $c(x,y)$.

Denote the left and right marginals of γ^i by ρ^i_+ and ρ^i_- for $i = 0, 1$. Define $\rho^i := \rho^i_+ - \rho^i_-$. From $\operatorname{spt}\gamma^1 \subset K \times K$ and $\operatorname{spt}\gamma^0 \subset \text{closure}(\tilde{K} \times \tilde{K})$ it is clear that ρ^1 vanishes outside K while ρ^0 vanishes inside K. Thus ρ^1 is the restriction of ρ to K, and ρ^0 its restriction to $\tilde{K}$, except that any Dirac mass concentrated on the end-points $a \in \operatorname{spt}\rho^i_+$ and $b \in \operatorname{spt}\rho^i_-$ may be divided between ρ^0 and ρ^1. Recall that $\rho \in \mathcal{M}_0^m(\mathbb{R})$ means $m+1$ intervals of supply are required to contain the full mass of ρ_+, where *interval of supply* refers to a maximal interval $L \subset \mathring{\mathbb{R}}$ satisfying $\rho_-[L] = 0$ but $\rho_+[L] > 0$. Label these intervals $I_0, I_1, \ldots, I_m$ around the circle $\mathring{\mathbb{R}}$, starting from $I_0 := I$ and continuing through K toward J. The interval J of demand lies between some pair I_j and I_{j+1}, where $1 \leqslant j \leqslant m-1$ since J does not lie adjacent to I. Thus the j intervals $I_1, \ldots, I_j$ in the interior of K, plus one interval containing I_0, represent the intervals of supply for ρ^1. Similarly, the $m - j$ intervals outside of K plus one interval containing I_0 represent the intervals of supply for ρ^0. Since the full mass of ρ^i_+ is contained in m (or fewer) intervals of supply, the inductive hypothesis will apply provided γ^i satisfies (i)–(iii). The first two conditions are met since $\operatorname{spt}\gamma^i \subset \operatorname{spt}\gamma$ and $\mathcal{C}(\gamma^i) \leqslant \mathcal{C}(\gamma) < \infty$. The third requires a more involved check.

Suppose γ^1 is compatible with an uncrossed network G_1 on the intervals of supply $I_0, I_1, \ldots, I_j$ and demand for ρ^1. Similarly, let γ^0 be compatible with an uncrossed network G_0 on the intervals of supply $I_{j+1}, \ldots, I_m, I_0$ and demand for ρ^0. (Such networks always exist in view of lemma 3.3). If the demand node for ρ^1 between I_j and I_0 is identified with the demand node for ρ^0 between I_0 and I_{j+1}, then ρ will be com-

patible with the uncrossed network G whose arcs are given by $\mathrm{arc}(G) := \mathrm{arc}(G_0) \cup \mathrm{arc}(G_1)$. Now suppose $\tilde\gamma^0 \in \Gamma_{G_0}(\rho^0_+, \rho^0_-)$ and $\tilde\gamma^1 \in \Gamma_{G_1}(\rho^1_+, \rho^1_-)$. Then $\tilde\gamma^0 + \tilde\gamma^1$ belongs to $\Gamma_G(\rho_+, \rho_-)$, so its transport cost can be compared to that of $\gamma = \gamma^0 + \gamma^1$ using (iii): $\mathcal{C}(\gamma^0 + \gamma^1) \leqslant \mathcal{C}(\tilde\gamma^0 + \tilde\gamma^1)$ or equivalently

$$\mathcal{C}(\gamma^0) + \mathcal{C}(\gamma^1) \leqslant \mathcal{C}(\tilde\gamma^0) + \mathcal{C}(\tilde\gamma^1).$$

Choosing $\tilde\gamma^{1-i} = \gamma^{1-i}$ and observing (ii) yields (iii) $\mathcal{C}(\gamma^i) \leqslant \mathcal{C}(\tilde\gamma^i)$ when $i = 0$, and similarly when $i = 1$. The inductive hypothesis can at last be applied to conclude optimality of γ^0 and γ^1.

Finally, Smith and Knott's criterion (theorem 5.3) shows $\operatorname{spt}\gamma^i$ to be c-cyclical monotone for $i = 0, 1$. Both supports include the point $(a, b) \in \mathbb{R}^2$, so applying proposition 5.5 to $S := \operatorname{spt}\gamma^1$ and $T := \operatorname{spt}\gamma^0 \setminus (K \times K)$ yields c-cyclical monotonicity of $S \cup T$. Since S has full measure for γ^1 while $T \cup \{(a, b)\}$ has full measure for γ^0, the union $S \cup T$ has full measure for $\gamma = \gamma^0 + \gamma^1$. Because $c(x, y)$ is continuous, taking closures of sets preserves c-cyclical monotonicity. In particular, the support of γ is contained in the closure of $S \cup T$ and therefore c-cyclically monotone. One last application of theorem 5.3 confirms that γ must be optimal. ■

6. Conclusions

The classical transportation result is the duality theorem that characterizes optimal distribution plans by the existence of a consistent shadow pricing scheme. The present study has had a different flavour, exploring the spatio-geometric implications of concave transportation costs (with respect to the distance). Concavity translates into an economy of scale for longer trips, which in certain situations will encourage cross-hauling. This tendency was shown to lead to the formation of a hierarchical structure in which price and distribution patterns repeat on different spatial scales. Specific predictions concerning qualitative and quantitative aspects of this structure were made that did not depend on details of the cost (e.g. the no-crossing rule and the rule of three), and therefore offer ready tests of the theory against empirical data.

The structure of the line was then exploited to show that, in one dimension, a complete solution to the problem does not require knowledge of the shadow prices everywhere, but merely the location of a few points where differentiability of this price must fail. These act like watersheds to separate different regions of the hierarchy, and can be located by an algorithmic sequence of finite-dimensional convex programs in standard form, each of which was represented as a network graphically. This reduction—from an infinite-dimensional problem with its continuous distribution of excess production to a handful of small network flow problems—required the sacrifice of linearity for convexity. It is nonetheless remarkable, and suggests that efficient distribution can generally be achieved by optimizing independently on different scales, and then allowing competition to adjust the boundaries between scales.

The author gratefully acknowledges the hospitality of the Institut des Hautes Etudes Scientifiques, as well as the support of an NSERC postdoctoral fellowship and grant DMS 9622997 of the NSF.

He offers warm thanks to Jochen Denzler, Christina Downs, Edson DeFaria, Stephen Semmes, Chris Shannon and John Stalker for many stimulating conversations. Vital remarks were contributed by Walter Craig and David Jerison, while Ivar Ekeland, Boris Mityagin, Stephen Semmes and Michael Reid pointed out respective connections with the work of Rochet, Bagdasarov, Thurston and Erdélyi & Etherington.

Appendix A. Non-decreasing maps and convex costs

For comparison's sake, we briefly recall the salient features of the transportation problem with strictly convex costs $c(x,y)=\ell(x-y)$ on the line: namely, that the optimal measure $\gamma\in\Gamma(\mu,\nu)$ must have non-decreasing support (definition 2.8), a restriction that characterizes this measure completely. These facts are well known and can be generalized to smooth costs satisfying the condition $\partial^2 c/\partial x\partial y<0$ used by Lorentz (1953), Spence and Mirrlees in the theories of rearrangements, signalling and optimal taxation (see Rochet (1987, § 3) for references).

Proposition A 1. *Let $c(x,y)=\ell(x-y)$ with $\ell\geqslant 0$ strictly convex on the line. If $\gamma\in\mathcal{M}_+(\mathbb{R}^2)$ is optimal then* $\operatorname{spt}\gamma\subset\mathbb{R}^2$ *must be non-decreasing.*

Proof. To produce a contradiction, suppose $\gamma\in\mathcal{M}_+(\mathbb{R}^2)$ is optimal yet fails to have non-decreasing support. Then there exist (x,y) and (x',y') in $\operatorname{spt}\gamma$ such that $(x'-x)(y'-y)<0$. Taking $x<x'$ costs no generality and forces $y'<y$. Thus both $x-y'$ and $x'-y$ must lie in the interval $(x-y,x'-y')$. Expressing $x-y'=(1-t)(x-y)+t(x'-y')$ yields $x'-y=t(x-y)+(1-t)(x'-y')$ with $t\in(0,1)$. Strict convexity of ℓ then gives

$$c(x,y')<(1-t)c(x,y)+tc(x',y'),$$
$$c(x',y)<tc(x,y)+(1-t)c(x',y').$$

But the sum of these two inequalities contradicts (2.1) and also theorem 5.3. ■

Proposition A 2. *Given measures $\mu,\nu\in\mathcal{M}_+(\mathbb{R})$ with the same total mass $\mu[\mathbb{R}]=\nu[\mathbb{R}]$, there is a unique joint measure $\gamma\in\Gamma(\mu,\nu)$ that has non-decreasing support.*

Proof. Existence of γ is simple: let $X,Y:[0,t]\longrightarrow\mathbb{R}$ be the non-decreasing maps defined on the interval of length $t:=\mu[\mathbb{R}]$ so that $\mu=X_{\#}\lambda_{[0,t]}$ and $\nu=Y_{\#}\lambda_{[0,t]}$. Setting $Z(\theta):=(X(\theta),Y(\theta))$, the measure $Z_{\#}\lambda_{[0,t]}$ has the correct marginals while its full mass lies on the non-decreasing curve $Z(\theta)$.

To address uniqueness, assume $\gamma\in\Gamma(\mu,\nu)$ has non-decreasing support. We shall show the values of γ to be completely determined by its marginals μ and ν. Since all Borel sets are generated by products $I_1\times J_1$ of semi-infinite intervals $I_1=[a,+\infty)$ and $J_1=[b,+\infty)$, it is enough to express $\gamma[I_1\times J_1]$ in terms of μ and ν.

Adopt the notation $I_0:=\mathbb{R}\setminus I_0$ and $J_0:=\mathbb{R}\setminus J_1$ for the complementary intervals, and define the two-by-two matrix $\gamma_{ij}:=\gamma[I_i\times J_j]$. Its column sums and row sums are prescribed to be $\mu_i:=\mu[I_i]$ and $\nu_j:=[J_j]$, for $i,j\in\{0,1\}$. These provide three independent constraints on the matrix γ_{ij} in terms of μ_0, ν_0 and $t=\mu_0+\mu_1=\nu_0+\nu_1$. There is a fourth constraint also: since $\operatorname{spt}\gamma$ is non-decreasing, either γ_{01} or γ_{10} must vanish. This leaves two possibilities only:

$$\begin{pmatrix}\gamma_{01}&\gamma_{11}\\ \gamma_{00}&\gamma_{10}\end{pmatrix}=\begin{pmatrix}0&t-\nu_0\\ \mu_0&\nu_0-\mu_0\end{pmatrix}\quad\text{or}\quad\begin{pmatrix}\mu_0-\nu_0&t-\mu_0\\ \nu_0&0\end{pmatrix}.\tag{A 1}$$

But one of these matrices has a negative entry, except in the special case $\mu_0=\nu_0$; either way, (A 1) selects $\gamma_{ij}\geqslant 0$ uniquely depending on $\mu[I_0]$ and $\nu[I_0]$. Since the intervals I_1 and J_1 were arbitrary, $\gamma\in\Gamma(\mu,\nu)$ is also unique. ■

Appendix B. Costs of concave type

This appendix explores some basic properties of the class of costs to which our results apply: the costs that we called *concave type* in definition 2.2. It begins with several alternative characterizations for this class of costs—sometimes subject to additional restrictions of smoothness or symmetry—and ends by bounding the transport cost $\mathcal{C}(\gamma)$ on $\Gamma(\rho_+, \rho_-)$, unless this cost turns out to be identically infinity.

The first lemma and its corollaries give a characterization of concave type in terms of monotonicity (see figure 8). They play a crucial role in propositions 4.2 and 5.5. This is followed by a lemma showing that additional symmetry reduces the costs $c(x, y)$ of concave type to strictly concave increasing functions of Euclidean distance $|x - y|$. A third lemma develops a differential characterization for smooth costs of concave type. It shows equivalence of concave type to the assertion that concavity of $c(x, y)$ in the direction leading away from the diagonal must outweigh convexity in the direction parallel to $y = x$.

We begin by observing that for $c : \mathbb{R}^2 \longrightarrow \mathbb{R} \cup \{-\infty\}$ of concave type, one knows $x \neq y$ and $y \neq z$ imply

$$c(x, y) + c(y, z) > c(x, z) + c(y, y). \tag{B 1}$$

To see this from the definition, set $x' = y$ and $y' = z$; the intersection of the circles $O(x, y)$ and $O(y, z)$ prevents (2.1) from being true. This version (B 1) of the strict triangle inequality asserts that, for a cost of concave type, it cannot be efficient to move mass from x to y and simultaneously from y to z; any factory at site y must be supplied from the on-site mine. Since the left side must be finite, we conclude as well that $c(x, y) > -\infty$ if $x \neq y$; infinite cost does not occur off the diagonal.

Lemma B 1. *For a cost $c : \mathbb{R}^2 \longrightarrow \mathbb{R} \cup \{-\infty\}$ to be of concave type, it is necessary and sufficient that $c(x, \cdot) - c(x', \cdot)$ increase strictly whenever $(\cdot)$ moves around the circle $\boldsymbol{S}^1 \approx \mathring{\mathbb{R}}$ from x toward $x' \neq x$.*

Proof. To establish necessity, assume the cost to be of concave type. Given four real numbers x, y', y and x', labelled in order (either clockwise or counterclockwise) around the circle $\mathbb{R} \approx \boldsymbol{S}^1$, we are asked to show

$$c(x, y') - c(x', y') < c(x, y) - c(x', y); \tag{B 2}$$

here $x = y'$ and $y = x'$ are allowed, but all other pairs must be distinct. The ordering of points forces the circles $O(x, y)$ and $O(x', y')$ to intersect, though $x \neq y$ and $y' \neq x'$. Since the cost is of concave type, one cannot have (2.1); the alternative (B 2) establishes the necessity claim. (Both sides of this inequality are unambiguously defined, since $c = -\infty$ does not occur off the diagonal.)

To prove the converse, suppose c fails to be of concave type. Then (2.1) is satisfied by four points $x \neq x'$ and $y \neq y'$, for which $O(x, y)$ and $O(x', y')$ intersect. These four points will be ordered x, y', y, x' on the unit circle, if not clockwise then counterclockwise. One also knows $x \neq y$, since otherwise the first circle degenerates yielding $x' = x = y$ or $x = y = y'$ (a contradiction). Similarly $x' \neq y'$. Strict monotonicity (B 2) of $c(x, \cdot) - c(x', \cdot)$ cannot hold without contradicting (2.1). This establishes the sufficiency claim. ∎

Corollary B 2. *An equivalent condition is that $c(\cdot, y) - c(\cdot, y')$ increases strictly whenever $(\cdot)$ moves around the circle $\boldsymbol{S}^1 \approx \mathring{\mathbb{R}}$ from y toward $y' \neq y$.*

Proof. From definition 2.2, $\tilde{c}(x,y) := c(y,x)$ will be of concave type whenever $c(x,y)$ is. The corollary follows by applying lemma B 1 to $\tilde{c}$. ■

Corollary B 3. *Fix a cost $c(x,y)$ of concave type. If four numbers x, y, x', y' remain ordered clockwise around the circle $\mathring{\mathbb{R}}$, the cross-difference $\Delta(x,y,x',y') := c(x,y) + c(x',y') - c(x,y') - c(x',y)$ will increase strictly when either y or y' is moved clockwise, or when x or x' moves counterclockwise, the other three arguments remaining fixed.*

Proof. The pairs of variables whose cost contributes to Δ with a positive sign will be called partners, e.g. $x \leftrightarrow y$, while the pairs which contribute with a negative sign will be called opposites, e.g. $x \leftrightarrow y'$. Fixing any three variables, e.g. y, x', y', the proposed motion slides the fourth variable (in this case x) away from its partner but toward its opposite. By lemma B 1 or its corollary, this increases the cross-difference Δ. ■

The next lemma couples with lemma 2.1 to show that for a cost invariant under translations and reflections of the line—$c(x,y) = c(x+z, y+z) = c(y,x)$—being of concave type is no different from being a strictly concave increasing function of the distance $|x-y|$.

Lemma B 4. *Suppose a cost $c : \mathbb{R}^2 \to \mathbb{R} \cup \{-\infty\}$ of concave type is invariant under $(x,y) \to (x+z, y+z)$ and $x \leftrightarrow y$. Then $c(x,y) = h(|x-y|)$ with h strictly concave increasing on $[0,\infty)$.*

Proof. Translation and reflection invariance implies $c(x,y) = h(|x-y|)$; the only issues at stake are the strict concavity and monotonicity of $h(x) = c(x,0)$ on $x \geqslant 0$. We begin by showing that h is increasing: $h(x'-y) < h(x'+y)$ for any $0 < y \leqslant x'$. Set $x = -x'$ and $y' = -y$ so that $x \leqslant y' < y \leqslant x'$. Then the circle $O(x,y)$ intersects $O(x',y')$ though the cost is of concave type. This precludes (2.1). Translation and reflection invariance yield $c(x,y) = c(x',y') = h(x'+y)$ and $c(x,y') = c(x',y) = h(x'-y)$, whence $2h(x'+y) > 2h(x'-y)$ as desired.

The next step is to prove h strictly midpoint concave, meaning $0 \leqslant 2x < 2x'$ must imply

$$2h(x+x') > h(2x) + h(2x'). \tag{B 3}$$

This time set $y := -x'$ and $y' = -x$ so that $y < y' \leqslant x < x'$. Again the circles $O(x,y)$ and $O(x',y')$ intersect, precluding (2.1). Using translation invariance to identify $c(x,y) = h(x+x') = c(x',y')$, $c(x,y') = h(2x)$ and $c(x',y) = h(2x')$, one recovers (B 3).

Having established strict midpoint concavity, the monotonicity provides sufficient smoothness to conclude concavity in the usual sense: a corresponding estimate holds for each convex combination of $2x$ and $2x'$. ■

The costs of concave type are next characterized by positivity off the diagonal of the mixed partial $\partial^2 c/\partial x \partial y$. Rotating coordinates by 45° reveals this condition to assert that the local concavity of $c(x,y)$ perpendicular to the diagonal must outweigh its convexity in the direction parallel to $y = x$. The limits expressed in (B 4) play the role of mixed partials at infinity, so this characterization is local on the torus $\mathring{\mathbb{R}} \times \mathring{\mathbb{R}}$. The proof is based on an observation (B 6) used by Rochet (1987) to discuss

costs satisfying the opposite condition $\partial c^2/\partial x\partial y < 0$ of Lorentz (1953), Spence and Mirrlees.

Lemma B 5. *Let $c(x,y)$ be continuous on the plane, and continuously twice differentiable on $\{(x,y) \mid x \neq y\} \subset \mathbb{R}^2$. For c to be a function of concave type, it is necessary that*

$$\frac{\partial^2 c}{\partial x \partial y} \geqslant 0 \text{ a.e.}, \qquad \lim_{x\to+\infty} \frac{\partial c}{\partial y}(-x,y) - \frac{\partial c}{\partial y}(x,y) \geqslant 0, \tag{B 4}$$

and

$$\lim_{y\to+\infty} \frac{\partial c}{\partial x}(x,-y) - \frac{\partial c}{\partial x}(x,y) \geqslant 0.$$

These three inequalities are also sufficient provided the first holds strictly (a.e.).

Proof. First assume $c(x,y)$ to be of concave type. For fixed $x < x'$, lemma B 1 shows $c(x,\cdot) - c(x',\cdot)$ to be strictly increasing as y increases from x toward x'. Thus

$$\frac{\partial c}{\partial y}(x,y) - \frac{\partial c}{\partial y}(x',y) \geqslant 0, \tag{B 5}$$

holds for $x < y < x'$. Letting $x' = -x$ tend to $+\infty$ yields the middle inequality (B 4). The last inequality must also hold true, since the condition for $c(x,y)$ to be of concave type is symmetrical in $x \leftrightarrow y$. (Existence of the limits can be deduced from non-negativity of the mixed partials by evaluating the integrand of (B 7).)

Now for the orderings $x < x' < y < y'$ or for $y < y' < x < x'$, observe

$$\int_y^{y'} \int_x^{x'} \frac{\partial^2 c}{\partial x \partial y} = c(x',y') - c(x',y) - c(x,y') + c(x,y). \tag{B 6}$$

Since the circles $O(x,y)$ and $O(x',y')$ intersect non-tangentially, (B 6) must be positive if the cost is to be of concave type. Taking the off-diagonal rectangle $[x,x']\times[y,y']$ sufficiently small forces the continuous function $\partial^2 c/\partial x\partial y \geqslant 0$ at $x \neq y$. This concludes the necessity proof.

To argue the converse, assume (B 4) holds with strict positivity a.e. of the mixed partial of the cost. We need to preclude (2.1) whenever the circles $O(x,y)$ and $O(x',y')$ intersect but $x \neq x'$ and $y \neq y'$. If this intersection takes place, by the $x \leftrightarrow y$ and *primed* $\leftrightarrow$ *unprimed* symmetries we may assume either (i) $x < x' \leqslant y < y'$ or (ii) $x \leqslant y' < y \leqslant x'$. In the first case, positivity of the integral (B 6) (letting y decrease to x' if necessary) precludes (2.1). In the second case, observe from (B 4) that

$$\begin{aligned} 0 < \int_{y'}^{y} \lim_{t\to+\infty} \left[\int_{-t}^{x} \frac{\partial^2 c}{\partial x \partial y}(u,v)\,\mathrm{d}u + \int_{x'}^{t} \frac{\partial^2 c}{\partial x \partial y}(u,v)\,\mathrm{d}u \right] \mathrm{d}v \\ \leqslant c(x,y) - c(x,y') - c(x',y) + c(x',y'), \end{aligned} \tag{B 7}$$

again contradicting (2.1). Thus c must be of concave type. ■

A final lemma is required to show that for $c(x,y) \geqslant 0$ of concave type, the moment conditions (4.4) control the transport cost $\mathcal{C}(\gamma)$ on $\Gamma_G(\rho_+,\rho_-)$.

Lemma B 6. *Fix $\rho \in \mathcal{M}_0(\mathbb{R})$, a Borel cost of concave type $c(x,y) \geqslant 0$, and $a,b,n,p,q \in \mathbb{R}$. Let $X,Y : [0,t] \longrightarrow \mathbb{R}$ be the non-decreasing maps representing $\rho_+ = X_\# \lambda_{[0,t]}$ and $\rho_- = Y_\# \lambda_{[0,t]}$, extended periodically with period $t = \rho_+[\mathbb{R}]$. Then (4.4) implies*

$$I := \int_a^b c(X(n+\theta), Y(p-\theta))\,\mathrm{d}\theta < \infty.$$

Proof. Setting $x = X(n+\theta)$, $z = Y(p-\theta)$ and $y = q$ in the triangle inequality (B 1) before integrating yields

$$\int_a^b c(X(n+\theta), q)\,\mathrm{d}\theta + \int_a^b c(q, Y(p-\theta))\,\mathrm{d}\theta \geqslant I + (b-a)c(q,q).$$

If the integration is over one full cycle, e.g. $a = 0$ and $b = t$, then (2.7) shows that the two terms on the left coincide with the finite integrals (4.4). Since the integrands are non-negative and periodic, both integrals must also converge for arbitrary $a, b \in \mathbb{R}$ thus proving finiteness of I. ■

For $\rho \in \mathcal{M}_0^m(\mathbb{R})$ and costs of concave type satisfying the monotonicity conditions

$$c(x,p) \leqslant c(x,y) \quad \text{and} \quad c(x,y) \geqslant c(p,y), \tag{B 8}$$

whenever $x \leqslant p \leqslant y$, a strong converse is true (though it will not be proven here): the moment conditions (4.4) are implied whenever $\mathcal{C}(\gamma) < \infty$ holds for a single measure $\gamma \in \Gamma(\rho_+, \rho_-)$. Even when a cost of concave type fails to satisfy (B 8), it can be modified to obtain a cost $\tilde{c}(x,y) = c(x,y) - f(x) - g(y)$ which does by subtracting

$$f(x) = \lim_{y \to +\infty} c(x,y) - c(0,y) \quad \text{and} \quad g(y) = \lim_{x \to +\infty} c(x,y) - c(x,0) \tag{B 9}$$

(the limits exist by lemma B 1). Moreover, $\tilde{c}(x,y)$ will share the optimal measures of $c(x,y)$. The moment conditions (4.4) therefore cause no loss in generality: the functional $\tilde{\mathcal{C}}(\gamma)$ is either bounded above or identically infinity. Should $\tilde{\mathcal{C}}(\gamma) := +\infty$, minimizing the 'renormalized' cost $\tilde{\mathcal{C}}_G(\phi)$ of (4.12) over feasible potentials, Φ_G, and networks, G, should still select the unique measures γ with c-cyclically monotone support from $\Gamma(\rho_+, \rho_-)$.

References

Bagdasarov, S. K. 1998 Kolmogorov–Landau problem and extremal Zolotarev ω-splines. *Dissertationes Math. (Rozprawy Mat.)* **379**, 1–81.

Beckmann, M. 1952 A continuous model of transportation. *Econometrica* **20**, 643–659.

Beckmann, M. J. & Puu, T. 1985 *Spatial economics: density, potential, and flow*. Studies in Regional Science and Urban Economics, vol. 14. Amsterdam: North-Holland.

Bertino, S. 1966 Una generalizzazione della dissomiglianza di C. Gini tra variabili casuali semplici. *Giornale dell'Istituto Italiano degli Attuari* **29**, 153–178.

Dall'Aglio, G. 1956 Sugli estremi dei momenti delle funzioni di ripartizione-doppia. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* **10**, 35–74.

Dantzig, G. B. 1951 Application of the simplex method to a transportation problem. In *Activity analysis of production and allocation* (ed. T. C. Koopmans), pp. 359–373. Cowles Commission for Research in Economics Monographs, vol. 13. New York: Wiley.

Erdélyi, A. & Etherington, I. M. H. 1940 Some problems of non-associative combinations (2). *Edinb. Math. Notes* **32**, 7–12.

Evans, L. C. & Gangbo, W. 1999 Differential equations methods for the Monge–Kantorovich mass transfer problem. *Mem. Am. Math. Soc.* **137**, 1–66.

Gangbo, W. & McCann, R. J. 1996 The geometry of optimal transportation. *Acta Math.* **177**, 113–161.

Hitchcock, F. L. 1941 The distribution of a product from several sources to numerous localities. *J. Math. Phys.* **20**, 224–230.

Hotelling, H. 1929 Stability in competition. *Econom. J.* **39**, 41–57.

Kantorovich, L. 1942 On the translocation of masses. *C. R. (Dokl.) Acad. Sci. URSS (N.S.)* **37**, 199–201.

Kellerer, H. G. 1984 Duality theorems for marginal problems. *Z. Wahrsch. Verw. Gebiete* **67**, 399–432.

Koopmans, T. C. 1949 Optimum utilization of the transportation system. *Econometrica* (Suppl.) **17**, 136–146.

Lieb, E. H. 1977 Existence and uniqueness of the minimizing solution of Choquard's nonlinear equation. *Stud. Appl. Math.* **57**, 93–105.

Lorentz, G. G. 1953 An inequality for rearrangements. *Am. Math. Mon.* **60**, 176–179.

Monge, G. 1781 Mémoire sur la théorie des déblais et de remblais. *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pp. 666–704.

Rachev, S. T. 1984 The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory Probab. Appl.* **29**, 647–676.

Rochet, J.-C. 1987 A necessary and sufficient condition for rationalizability in a quasi-linear context. *J. Math. Econom.* **16**, 191–200.

Rockafellar, R. T. 1984 *Network flows and monotropic optimization*. New York: Wiley.

Rüschendorf, L. 1991 Fréchet bounds and their applications. In *Advances in probability distributions with given marginals* (ed. G. Dall'Aglio, S. Kotz & G. Salinetti), pp. 151–187. Mathematics and its applications, vol. 67. Dordrecht: Kluwer.

Smith, C. & Knott, M. 1992 On Hoeffding–Fréchet bounds and cyclic monotone relations. *J. Multivariate Analysis* **40**, 328–334.

Tchen, A. H. 1980 Inequalities for distributions with given marginals. *Ann. Probab.* **8**, 814–827.

Thurston, W. P. 1986 Earthquakes in two-dimensional hyperbolic geometry. In *Low-dimensional topology and Kleinian groups* (ed. D. B. A. Epstein), pp. 91–112. London Mathematical Society Lecture Note Series, vol. 112. Cambridge University Press.

Uckelmann, L. 1997 Optimal couplings between one-dimensional distributions. *Distributions with given marginals and moment problems* (ed. V. Benes & J. Stepan), pp. 275–281. Dordrecht: Kluwer.