Optimal Transportation: Continuous and Discrete

Master Project under the supervision of

Professor Victor Panaretos

Yoav Zemel

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Chair of Mathematical Statistics
Mathematics Section

June 22, 2012

ABSTRACT

This Master Project presents some basic aspects of the Monge–Kantorovich problem, also known in the literature as the "mass transportation problem". There are two goals for this project. The first is to survey theoretical ideas, such as existence and uniqueness proofs and characterization of the solutions. The second is to deal with the practical question of explicitly finding the solutions for two families of problems: the discrete ones and the absolutely continuous ones. In the introduction, the problem is formulated in measure theoretical terms as an optimization problem over measures, and examples are given. In Section 2 a dual problem for the Monge–Kantorovich problem, involving optimization over functions, is surveyed and a duality theorem is given. In the particular case where a finite number of points is to be transported to equally many locations, the problem can be recast as a linear program; the foundations of the latter is the topic of Section 3. In Section 4 an efficient algorithm for this linear program, the Hungarian method, is presented. Section 5 deals with absolutely continuous measures in Euclidean spaces. The particular cost function, squared Euclidean distance, gives rise to an algorithm for finding the optimal transportation; this is sketched in Section 6. In the last section, we return to the case of arbitrary measures, and optimality is characterized by a property called c-monotonicity. An elegant stability result of the Monge–Kantorovich problem is established as a corollary. This is seen to provide, also, a heuristic for interpreting transportation of absolutely continuous measures via the Hungarian method.

The image to the left in the title page is from `http://de.wikipedia.org/w/index.php?title=Datei:SBB_482_Kesselwagen.jpg&filetimestamp=20090717124221` and is attributed to the Wikipedia user Niesen. The second image is from `http://upload.wikimedia.org/wikipedia/commons/e/e6/Lego-Chicago_Rush_Hour.jpg`, and is attributed to the user Joe. Both were retrieved June 22, 2012.

Contents

Notation

When X is a topological space, we treat it as a measurable space with its Borel field, the smallest σ-algebra containing the open sets of X. $M(X)$ is the set of finite signed measures on X, $M_+(X)$ is the set of positive finite measures on X and $P(X)$ is the set of probability measures on X. For a measure $\mu \in M_+(X)$, its support $supp(\mu)$ is the smallest closed set $A \subseteq X$ for which $\mu(A) = \mu(X)$. It is well-defined at least when X is a Polish space, that is, a complete separable metric space. For a set E, 1_E is the indicator function of E, which we sometimes denote by $1E$. The dirac measure on x is the measure δ_x defined by $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. It will sometimes be denoted by $\delta\{x\}$. When $\nu \in P(X)$ and $\mu \in P(Y)$, $\Pi(\mu, \nu)$ is the set of couplings μ and ν. These are measures π on $X \times Y$ such that $\pi(X \times B) = \nu(B)$ and $\pi(A \times Y) = \mu(A)$ for any measurable A and B.

The inequality $y \geq x$ means that x and y are vectors of the same dimension d, and $y_i \geq x_i$ for $i = 1, \ldots, d$. $\mathbb{R}$ is the set of real numbers, $\mathbb{R}_+$ is the set of positive real numbers and when d is an integer, $\mathbb{R}_+^d = [0, \infty)^d$. For an integer n, $[n] = \{1, \ldots, n\}$, and $[0]$ is the empty set. S_n is the set of permutations on $[n]$, that is bijective functions from $[n]$ to itself.

$M_{m,k}$ is the set of $m \times k$ real matrices. For such a matrix A, $A_i \in \mathbb{R}^m$ is the i-th column of A, and $A^i \in \mathbb{R}^k$ is the i-th row of A. $M_n = M_{n,n}$ is the set of square matrices of order n, and $I_n \in M_n$ is the identity matrix.

$\binom{n}{k} = n!/(k!(n-k)!)$ is the binomial coefficient of n and k; this is the coefficient of x^k in the polynomial $(1+x)^n$.

1 Introduction

1.1 The Monge and Kantorovich problems

Imagine that your child has built a structure of volume A from Lego pieces in the living room of your house. Having invited guests for dinner, you ask your child to remove his structure so that you could comfortably host your guests in the living room. Your child asks "but where should I move it to?". You stare at the structure for a moment, then remember that there is box in his room. Miraculously, the box has exactly the same volume A. You tell your child to put the pieces in the box. Your child has always dreamed of being an optimizer; he tries to think how to move his creation with a minimal effort.

Monge [10] tried to deal with this sort of questions (since Lego did not exist in the 18th century, one can say for certain that the preceding paragraph has not been Monge's motivation). His **transport problem** can be formulated in measure theoretic terms as follows. Given measurable spaces (X, μ) ("the structure") and (Y, ν) ("the box"), $\mu(A)$ and $\nu(B)$ represent the mass of the (measurable) subsets $A \subseteq X$ and $B \subseteq Y$. Of

course, transporting $x \in X$ into $y \in Y$ requires some work, or effort, or cost $c(x, y)$. $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ is called the **cost function**. It is assumed to be measurable and in most cases nonnegative, and may take infinite values.
We wish to transport the elements of X into Y, in a way that will preserve the mass. A measurable $T : X \to Y$ can only be a reasonable candidate if for any measurable $B \subseteq Y$, $\mu(T^{-1}(B)) = \nu(B)$. In words, the mass transported into B has to equal the mass of B. For such T, we say that ν is the **push-forward** of μ by T, and write

$$\nu = T\#\mu.$$

Such T will be called **transport map**.
Letting $B = Y$, we obtain an obvious necessary condition for the existence of transport maps: it must be that $\mu(X) = \nu(Y)$, and we will always assume that this quantity is finite. The measures μ and ν can then be normalized to become probability measures.
The Monge problem is thus: find a transport map $T : X \to Y$ that minimizes

$$I(T) \stackrel{\text{def}}{=} \int_X c(x, T(x)) d\mu(x).$$

The Kantorovich problem

If $T : X \to Y$ is a transport map, it maps $x \in X$ to some $T(x) \in Y$. If $\{x\}$ has a positive mass, then this mass has to be moved entirely to $T(x)$. Kantorovich [7] proposed a relaxation to the Monge problem by allowing mass to be split. Informally, for any $x \in X$ one constructs a probability measure $\pi_x \in P(Y)$ which describes how the mass of x is distributed in Y. Formally, we consider measures π on the product space $X \times Y$ with marginal distributions μ on X and ν on Y, that is

$$\pi(A \times Y) = \mu(A) \quad \forall A \subseteq X \text{ measurable}, \tag{1.1}$$

and

$$\pi(X \times B) = \nu(B) \quad \forall B \subseteq Y \text{ measurable}. \tag{1.2}$$

(1.1) means that the mass transported from A, $\pi(A \times Y)$, equals the mass of A. (1.2) means that the mass transported into B, $\pi(X \times B)$, equals the mass of B. If π satisfies these two requirements, it is called **transference plan**.
For example, if ν is the push-forward of μ by T, we can define a transference plan $\pi = (id \times T)\#\mu$ by

$$\pi(A \times B) = \mu\left(\{x \in A : T(x) \in B\}\right) = \mu(A \cap T^{-1}(B)), \tag{1.3}$$

so π gives no measure to subsets of $X \times Y$ that are disjoint to the set $\{(x, T(x))\}_{x \in X}$. We sometimes write (1.3) in a more compact and somewhat sloppy notation,

$$d\pi(x, y) = d\mu(x) 1\{y = T(x)\},$$

where $1\{\cdot\}$ is the indicator function. Equivalently, for any continuous and bounded $\varphi : X \to \mathbb{R}$,

$$\int_{X\times Y} \varphi d\pi = \int_X \varphi(x, T(x)) d\mu(x).$$

Therefore a transport map T always induces a transference plan π.
It is convenient to denote the set of transference plans by

$$\Pi(\mu,\nu) \stackrel{\text{def}}{=} \{\pi \in P(X\times Y) : \forall A \;\; \pi(A\times Y) = \mu(A) \quad \text{and} \quad \forall B \;\; \pi(X\times B) = \nu(B)\},$$

where $A \subseteq X$ and $B \subseteq Y$ are measurable sets.
The **Kantorovich problem** is to find

$$\inf_{\pi\in\Pi(\mu,\nu)} I(\pi), \qquad I(\pi) \stackrel{\text{def}}{=} \int_{X\times Y} c(x,y) d\pi(x,y). \tag{1.4}$$

The Kantorovich problem is a relaxation of the Monge problem, since any transference map induces a transference plan that allocates I the same value. In many senses, the Kantorovich problem is easier to handle than the Monge problem. For example, transference plans always exist: simply consider the measure $\pi(A\times B) = \mu(A)\nu(B)$. Transport maps, however, need not exist in general. Furthermore, under fairly weak conditions the Kantorovich problem admits a minimizer; stronger conditions are needed to guarantee the existence of transport maps. $\Pi(\mu,\nu)$ is a convex set and $I(\pi)$ is linear with π, while the constraint $\nu = T\#\mu$ is "nonlinear" with T. The Kantorovich problem also introduces symmetry between μ and ν: given a transference plan π from μ to ν, $\tilde{\pi}(B\times A) = \pi(A\times B)$ is a transference plan from ν to μ. In this sense, the Monge problem can be asymmetric.

Since a transport map is (sloppily speaking) a transference plan, it is clear that the value obtained from the Kantorovich problem is at least as small as the one obtained from the Monge problem. A natural question is when these values coincide, and we provide some examples for conditions guaranteeing equality.

We would also like to briefly discuss a probabilistic view of these problems. Let U and V be random elements from some probability space $(\Omega, \mathcal{F}, P)$ taking values in X and Y respectively and having laws μ and ν respectively:

$$P(V^{-1}(B)) = \nu(B) \qquad \text{and} \qquad P(U^{-1}(A)) = \mu(A).$$

(So $\nu = V\#P$ and $\mu = U\#P$.) A **coupling** of U and V is a random element Z from Ω, with values $X\times Y$, whose law has the laws of U and V as marginals. This means that $P(Z^{-1}(A\times Y)) = P(U^{-1}(A)) = \mu(A)$ and $P(Z^{-1}(X\times B)) = P(V^{-1}(B)) = \nu(B)$ for all measurable $A \subseteq X$, $B \subseteq Y$. The Kantorovich problem is then to minimize the expected value

$$Ec(Z)$$

over all couplings of U and V. In other words, we are free to change U and V and the underlying probability space, as long as the constraints $\nu = V\#P$ and $\mu = U\#P$ are maintained. At least one such coupling always exists, since a construction making U and V independent is always possible. In order to solve the Monge problem, we only consider deterministic couplings, in the sense that $V = T(U)$ is a deterministic function of U. Such functions need not exist in general. Also, the existence of such T does not imply the existence of a deterministic function S such that $S(V) = U$; the Monge problem is asymmetric.

Before proceeding with a few examples, we provide a simple but useful characterization of $\Pi(\mu, \nu)$ with respect to measurable functions. The proof, immediate from the definitions, is omitted.

Proposition 1.1 *$\pi \in \Pi(\mu, \nu)$ if and only if for any $(\varphi, \psi) \in L_1(\mu) \times L_1(\nu)$ we have*

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

1.2 Examples

We remind the reader that a dirac measure on x is the measure

$$\delta\{x\}(A) = \delta_x(A) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise.} \end{cases}$$

Example 0: $\mu = \delta_x$, but $\nu \neq \delta_y$ for any y. It is impossible to map a dirac measure to a non dirac measure: if $T : X \to Y$, then $\nu(\{T(x)\}) < 1 = \mu(T^{-1}(\{T(x)\}))$; No transport maps exist. The unique transference plan is the independent coupling: distribute x on Y according to ν, i.e. $\pi = \delta_x \otimes \nu = \mu \otimes \nu$. This costs

$$\int_{X\times Y} cd\pi = \int_Y c(x, y)d\nu(y).$$

If μ and ν change roles ($\nu = \delta_y$), then the unique transference plan is the unique transport map that sends any $x \in X$ to y.

Example 1: Let λ be Lebesgue measure on $\mathbb{R}$, $X = Y = [0, n + 1]$, $c(x, y) = |x - y|^p$ ($p > 0$), $\mu = \frac{1}{n}\lambda|_{[0,n]}$ and $\nu = \frac{1}{n}\lambda|_{[1,n+1]}$. Consider the transport maps

$$T_1(x) = x + 1 \qquad \text{and} \qquad T_2(x) = \begin{cases} x & 1 < x \leq n \\ x + n & 0 \leq x \leq 1. \end{cases}$$

T_2 leaves all the common mass in place, while T_1 does not. One readily calculates

$$I(T_1) = \int_X |x - T_1(x)|^p d\mu(x) = 1, \qquad I(T_2) = \int_X |x - T_2(x)|^p d\mu(x) = n^{p-1}.$$

We see that T_1 yields a better solution if and only if $p > 1$, and if $p = 1$ both are equally good. This is not surprising, since $x \to x^p$ is a concave function for $p < 1$ and convex for $p > 1$. Consider the case $p = 1$, where both T_1 and T_2 incur a cost of 1.
Let φ be a 1-Lipschitz function ($|\varphi(x) - \varphi(y)| \leq |x - y|$ for any x, y) and T be some transport map (not necessarily optimal). Then

$$\int_X \varphi d(\mu - \nu) = \int_X (\varphi(x) - \varphi(T(x)))\, d\mu(x) \leq \int_X |x - T(x)| d\mu(x).$$

Therefore, if we find a 1-Lipschitz φ and a transport plan T for which equality holds, then T must be a minimizer for I (and φ must be a maximizer for $\int_X \varphi d(\mu - \nu)$). Choosing $\varphi(x) = -x$, we have

$$\int_X \varphi d(\mu - \nu) = \int_X x d\nu(x) - \int_X x d\mu(x) = \frac{1}{n}\int_1^{n+1} x dx - \frac{1}{n}\int_0^n x dx = \frac{(n+1)^2 - 1^2 - n^2}{2n} = 1.$$

It follows that 1 is the minimal value that can be obtained for I by transport maps, so both T_1 and T_2 are optimal. Furthermore, it turns out that this Lipschitz test is also valid for transference plans; T_1 and T_2 are thus optimal as transference plans too. It will be shown that if X and Y are metric spaces with the metric c, then a solution to the Kantorovich problem that leaves all common mass in place exists (see Subsection 2.3).

Example 2: $\mu = \frac{1}{3}\delta\{x_1\} + \frac{2}{3}\delta\{x_2\}$, $\nu = \frac{1}{3}\delta\{y_1\} + \frac{2}{3}\delta\{y_2\}$ with $x_1 \neq x_2$. Let $c_{ij} = c(x_i, y_j)$. The unique transport map is $T(x_i) = y_i$, and costs $c_{11}/3 + 2c_{22}/3$. Suppose that we send a_{ij} mass from x_i to y_j. This will cost $\sum_{i,j} a_{ij}c_{ij}$. A measure $\pi = \sum_{i,j} a_{ij}\delta\{(x_i, y_j)\}$ is an admissible transference plan if and only if the following constraints are satisfied:

$$\begin{array}{ll} 0 \leq a_{11} \leq 1/3 & a_{11} + a_{12} = 1/3 \\ 0 \leq a_{12} \leq 1/3 & a_{12} + a_{22} = 2/3 \\ 0 \leq a_{21} \leq 2/3 & a_{11} + a_{21} = 1/3 \\ 0 \leq a_{22} \leq 2/3 & a_{21} + a_{22} = 2/3 \end{array}$$

Equivalently,

$$a_{12} = a_{21} = 1/3 - a_{11}, \quad a_{22} = 1/3 + a_{11}, \qquad 0 \leq a_{11} \leq \frac{1}{3},$$

where the choice of a_{11} incurs the cost

$$I(a_{11}) = \frac{c_{12} + c_{21} + c_{22}}{3} + a_{11}(c_{11} + c_{22} - c_{12} - c_{21}).$$

This is a linear function of a_{11}, so the optimal solution is either $a_{11} = 0$ or $a_{11} = 1/3$ (or any a_{11}), depending on the sign of $c_{11} + c_{22} - c_{12} - c_{21}$. The unique transport map corresponds to $a_{11} = 1/3$, while $a_{11} = 0$ means sending x_1 to y_2 and then evenly splitting x_2 between y_1 and y_2. It is easy to set up the details so that $a_{11} = 0$ be the unique transference plan, in which case both the Monge problem and the Kantorovich problem have a unique solution, but the values obtained from the two problems differ.

Observe also another difficulty of the Monge problem—if we slightly perturb μ to equal $(1/3-\varepsilon)\delta\{x_1\}+(2/3+\varepsilon)\delta\{x_2\}$, then transport maps no longer exist. The value of the Kantorovich problem, however, changes only by $O(\varepsilon)$.

Example 3 (**the discrete Monge–Kantorovich problem**): $\mu=(1/n)\sum_{i=1}^n \delta\{x_i\}$ and $\nu=(1/n)\sum_{j=1}^n \delta\{y_j\}$. Transport maps here involve mapping x_i to $y_{\sigma(i)}$ for some permutation σ. Thus, the Monge problem is to find a permutation $\tau\in S_n$ that minimizes

$$I(\tau)=\frac{1}{n}\sum_{i=1}^n c\left(x_i,y_{\tau(i)}\right).$$

For the Kantorovich problem, suppose that we send $a_{ij}\geq 0$ mass from x_i to y_j, then transference plans are of the form $\sum_{i,j} a_{ij}\delta\{(x_i,y_j)\}$. Since x_i has to be sent entirely, we must have $\sum_{j=1}^n a_{ij}=1/n$ and since the total mass sent to y_j must equal its mass, we must also have $\sum_{i=1}^n a_{ij}=1/n$. The cost of a transference plan is $\sum_{i,j} a_{ij}c_{ij}$ where $c_{ij}=c(x_i,y_j)$. Replacing a_{ij} by $b_{ij}=na_{ij}$ we see that transference plans are equivalent to **bistochastic matrices**

$$B_n=\left\{B\in\mathbb{R}_+^{n^2}:\forall i\ \sum_{j=1}^n b_{ij}=1\quad\text{and}\quad\forall j\ \sum_{i=1}^n b_{ij}=1\right\}. \tag{1.5}$$

The Kantorovich problem is to find a matrix $B=(b_{ij})\in B_n$ that minimizes

$$I(B)=\frac{1}{n}\sum_{i,j=1}^n b_{ij}c_{ij},\qquad B\in B_n.$$

$B\in B_n$ is a **permutation matrix** if $b_{ij}\in\{0,1\}$, in which case any row has exactly one entry that equals 1, and any column has exactly one entry that equals 1. The Monge problem is to find a minimizer which is also a permutation matrix.
Since B_n is a convex set and I is linear, we know that there must be an extreme point of B_n that minimizes I (an extreme point in a convex set is a point that cannot be written as convex combination of other points in the set). By Birkhoff's theorem, the extreme points of B_n are exactly the permutation matrices. Therefore, in this case, the set of optimizers for the Kantorovich problem is precisely the convex-hull of the set of optimizers for the Monge problem. The corresponding values are the same, and uniqueness in one of the problems is equivalent to uniqueness in the other.
Observe that the constraints (1.5) on B as well as the cost $\sum_{i,j} b_{ij}c_{ij}$ are all linear with respect to the variables $(b_{ij})_{i,j=1}^n$ and the Kantorovich problem can be formulated as a linear program; this is the topic of Section 3. An efficient algorithm of complexity bounded by $O\left(n^4\right)$ is presented in Section 4.

Example 4: Let $Q=[0,1]^{n-1}$, $X=Q\times\{0\}$, $Y=Q\times\{-1,1\}$. Let μ,ν be the uniform probability measures on X and Y respectively and $c=h(\|x-y\|)$ for a continuous, strict

monotone h with $h(1) = 1$. Since $c(x, y) \geq 1$, $I(\pi) \geq 1$ for any transference plan π. Therefore, $I(\pi) = 1$ if and only if π is concentrated on the set

$$c^{-1}(\{1\}) = \{(x, y) \in X \times Y : c(x, y) = 1\} = \{(x, y) \in X \times Y : \|x - y\| = 1\}.$$

If $u \in Q$, splitting $(u, 0)$ evenly among $(u, 1)$ and $(u, -1)$ will provide such a transference plan: let

$$d\pi(x, y) = d\mu(x) \left(\frac{1}{2} 1\{y = x + \varepsilon_n\} + \frac{1}{2} 1\{y = x - \varepsilon_n\} \right), \qquad \varepsilon_n = (0, 0, 0, \ldots, 1) \in \mathbb{R}^n.$$

Then π is supported on $c^{-1}(\{1\})$ and is a transference plan with $I(\pi) = 1$. It is not difficult to see that π is unique; the formal details are skipped for brevity. Since π is not induced from any transport map, we can conclude that $I(T) > 1$ for any transport map T. However, we can approach one as much as we wish, so $\inf I(T) = 1$.
To see this, assume for simplicity $n = 2$. We approximate the optimal π as follows. Let $k \in \mathbb{N}$ and map the segment $[0, 1/(2k)) \times \{0\}$ to $[0, 1/k) \times \{1\}$, then the segment $[1/(2k), 1/k) \times \{0\}$ to $[0, 1/k) \times \{-1\}$, and continue similarly. This can be done with the piecewise linear function

$$T_k(x, 0) = \begin{cases} \left(2x - \frac{i}{k}, 1\right) & x \in [\frac{2i}{2k}, \frac{2i+1}{2k}) \\ \left(2x - \frac{i+1}{k}, -1\right) & x \in [\frac{2i+1}{2k}, \frac{2i+2}{2k}) \\ (1, 1) & x = 1. \end{cases}$$

It is easy to see that T_k is a transport map, and by continuity and monotonicity of h,

$$I(T_k) \leq h\left(\left\| \left(\frac{1}{2k}, 1 \right) \right\| \right) \to 1, \qquad k \to \infty.$$

A similar sequence of functions can be constructed when $n > 2$.
To conclude, the values obtained from the two problems are equal, but only the Kantorovich problem admits a solution.

Example 5: $X = [0, 1]$, $Y = [1, 2]$ with Lebesgue measure and $c(x, y) = |x - y|^p$, $p > 0$ (this is example 1 with $n = 1$). Here again the optimal transport map depend on whether $p > 1$ or $p < 1$. Reasonable candidates are

$$T_1(x) = x + 1 \qquad \text{and} \qquad T_2(x) = 2 - x,$$

with costs

$$c(T_1) = \int_0^1 1 dx = 1 \qquad \text{and} \qquad \int_0^1 (2 - 2x)^p dx = \frac{2^p}{p + 1}.$$

If $p > 1$, T_1 is better; if $p < 1$, T_2 is better. If $p = 1$ both incur a cost of one. In fact, the case $p = 1$ is completely degenerate because the cost function splits into functions of one variable only: $c(x, y) = y - x$. If π is a transference plan, then

$$I(\pi) = \int_{X \times Y} (y - x) d\pi(x, y) = \int_Y y dy - \int_X x dx = \frac{2^2 - 1^2 - 1^2}{2} = 1.$$

All transference plans (and therefore all transport maps) give the same value.
Observe that $p < 1$ yields a concave cost function, so that "one long trip and one short trip are better than two medium trips". For this reason the optimal map T_2 reverses the orientation. When $p > 1$ the cost function is convex and long trips are avoided, therefore T_1 is the optimal map. Gangbo and McCann [6] show that these maps are the unique optimal *transference plans* and describe the geometric properties of the optimal maps for these and more general cost functions (see also Subsection 5.2).

1.3 Outline

This project is organized as follows. In the next subsection we state several elementary results about measures on Polish spaces and lower semi-continuous functions. A good reference for the former is the book by Billingsley [2]. As for lower semi-continuity, we will only use Proposition 1.4 and its converse, Lemma 1.5. Thus no reference will be needed.
Many optimization problems admit **dual formulations**; in Section 2 we present the dual problem of the Monge–Kantorovich problem. While the original problem (**primal problem**) is to minimize an integral with respect to *measures*, the dual problem consists of maximizing integrals with respect to *functions*. The main result of this section is Theorem 2.3, stating that the values obtained from both problems are equal. Then the above examples will be reviewed (and solved!) in the context of duality. Finally, the particular form of the dual problem when the cost function is a metric will be proved, and a corollary about leaving the common mass in place will follow.
The next two sections are devoted to Example 3 above. The example is recast as a linear program, and the foundations of linear programming are surveyed in Section 3. The notion of basic solutions will be introduced, and the simplex algorithm will be presented. Geometrical interpretations, as well as the relation to convex sets, will be sketched. The section is concluded by an example illustrating the limitations of the simplex method.
In Section 4 we take advantage of the special form of the discrete Monge–Kantorovich problem as a linear program. This section describes the Hungarian algorithm, or the Hungarian method, that solves the problem efficiently. Due to the structure of the problem, this section is combinatorial in nature, and relations to graph theory will be established. However, these are not crucial for the understanding of the algorithm, and if the reader is not interested he can safely skip the graph-theoretic parts.
When the measures are on the Euclidean space $\mathbb{R}^d$ and are absolutely continuous with respect to Lebesgue measure, a special relation holds between the solution of the primal problem and the solution of its dual. This relation, as well as geometrical properties of the solutions, is discussed in Section 5. We will mainly treat the quadratic cost function $c(x, y) = \|x - y\|^2$, where the notion of convexity plays a key role. Lastly, we state analogous results for more general cost functions.

In Section 6 we again restrict attention to absolutely continuous measures, supported on compact sets in $\mathbb{R}^d$ with quadratic cost function. In this particular case we recast the dual problem as an unconstrained optimization problem of one function φ. We show that the objective function is differentiable (in the appropriate sense) with respect to the unknown function φ. This result gives rise to a gradient descent algorithm in order to find the optimal solution to the dual problem and, using the relation established in Section 5, the optimal transport map.
In Section 7 the setting is once again very general. Notions from Section 5 are generalized, which allows for a characterization of optimal transference plans by a property of their support, c-monotonicity. An elegant result regarding the stability of the Kantorovich problem with respect to weak convergence is established. It is also an illustration of how, in some sense, the discrete Monge–Kantorovich problem approximates the general problem.

1.4 Several elementary results

We always assume that X and Y are separable metric spaces (**Polish spaces**). Measures on X and Y will always be defined on their **Borel σ-algebra**, the smallest σ-algebra containing all the open sets. We denote by $P(X)$, $M_+(X)$ and $M(X)$ the set of probability measures, positive (finite) measures and signed measures on X. The reason we use Polish spaces is that any measure on such spaces is tight:

Definition 1.2 *A set of probability measures $\{\mu_a\}$ is **tight** if for any ε there exists a compact set K such that $\mu_a(K) > 1 - \varepsilon$ for any a.*

An arbitrary measure on an arbitrary metric space may fail to be tight: take, for example, X to be an uncountable discrete space and $\mu(A)$ to equal 1 if $X \setminus A$ is uncountable and 0 otherwise. Compact sets are finite and therefore measure-zero, while the space itself has measure 1. When X is Polish, these anomalies no longer exist—any measure on a Polish space is tight [2, Theorem 1.3].
We say that a sequence of measures $\{\mu_k\} \subseteq P(X)$ **converges weakly** to $\mu \in P(X)$, denoted by $\mu_k \xrightarrow{\mathrm{D}} \mu$, if $\mu_k(A) \to \mu(A)$ for any measurable A such that $\mu(\partial A) = 0$. ∂A is the **boundary** of A, the set of points in the closure of A that are not in its interior. In fact, $P(X)$ is metrizable by weak convergence: the **Lévy–Prokhorov metric** between μ_k and μ converges to 0 if and only if $\mu_k \xrightarrow{\mathrm{D}} \mu$. Furthermore, when X is Polish, so is $P(X)$. By the Portmanteau theorem [2, Theorem 2.1], $\mu_k \xrightarrow{\mathrm{D}} \mu$ if and only if for any continuous bounded function $\varphi : X \to \mathbb{R}$, $\int_X \varphi d\mu_k \to \int_X \varphi d\mu$ when k tends to infinity. The collection of continuous bounded functions approximates any integrable function in the sense that for any $g \in L_1(\mu)$ there exists a continuous bounded f such that $\int_X |f - g| d\mu < \varepsilon$ for arbitrary $\varepsilon > 0$.

We will usually deal with continuous cost functions. Nevertheless, for many results lower semi-continuity is sufficient.

Definition 1.3 *A real-valued function (possibly taking infinite values) f on a metric space (X, d_X) is* ***lower semi-continuous*** *at $x \in X$ if*

$$f(x) \leq \liminf_{y \to x} f(y).$$

f is lower semi-continuous if it is lower semi-continuous at any $x \in X$.

If f is lower semi-continuous at x but *not* continuous there, then $f(x)$ is "too small" and not too large; the function $f(x) = 1\{x \neq 0\}$ is lower semi-continuous at 0 while $1 - f(x) = 1\{x = 0\}$ is not. It is easy to see that f is lower semi-continuous if and only if $\{x : f(x) \leq c\}$ is a closed set (so f is measurable). Also, f is continuous if and only if both f and $-f$ are lower semi-continuous. Another useful characterization of lower semi-continuity is being a monotone limit of continuous functions.

Proposition 1.4 *Let f be lower semi-continuous and nonnegative. Then there exists an increasing sequence of continuous bounded functions that converge pointwise to f. Consequently, lower semi-continuous functions are Borel measurable.*

Proof. Let d denote the metric on of the space X. Define

$$f_n(x) = \inf_{y \in X} \left(f(y) + nd(x, y) \right).$$

It is clear that $f_n(x) \leq f_{n+1}(x)$, and letting $y = x$ gives $f_n(x) \leq f(x)$. f_n is n-Lipschitz, so it is uniformly continuous.
For convergence, we split into cases. Suppose that $f(x)$ is infinite. For $M \in \mathbb{R}$, we show that $\liminf_{n \to \infty} f_n(x) \geq M$. Lower semi-continuity of f implies the existence of $\varepsilon > 0$ such that $f(y) \geq M$ if $d(x, y) \leq \varepsilon$, so $f_n(x) \geq \min(M, n\varepsilon)$ and the assertion follows.
If $f(x)$ is finite, let x_n such that $f_n(x) \geq f(x_n) + nd(x, x_n) - 1/n$. As f is nonnegative this means $nd(x, x_n) \leq f_n(x) + 1/n \leq f(x) + 1/n < \infty$, so $x_n \to x$. By nonnegativity of the metric and lower semi-continuity, we obtain

$$\liminf_{n \to \infty} f_n(x) \geq \liminf_{n \to \infty} f(x_n) - 1/n \geq f(x),$$

establishing $f_n(x) \nearrow f(x)$ for any $x \in X$. Replacing f_n by $\min(n, f_n)$ proves the proposition.
Proposition 1.4 is a characterization of lower semi-continuity, because its converse also holds (the proof is immediate from the definitions).

Lemma 1.5 *Let A be an arbitrary set and f_a lower semi-continuous for any a. Then*

$$f(x) \stackrel{\text{def}}{=} \sup_{a \in A} f_a(x)$$

is lower semi-continuous. If f_a is convex for any a then so is f.

2 The Kantorovich duality

2.1 The duality theorem

In some cases, optimization across functions is simpler than optimization across measures. Kantorovich discovered a dual problem to the minimization of $I(\pi)$ over the transference plans between μ and ν. This is the topic of this section. From now on, we assume that the cost function c is nonnegative.

As mentioned in the introduction, an advantage of the Kantorovich problem over the Monge problem is that transference plans always exist. When working on Polish spaces with lower semi-continuous functions, a little bit of work shows that an *optimal* transference exists. When μ and ν are probability measures on spaces X and Y, we remind the reader that $\Pi(\mu, \nu)$ is the set of probability measures on the product space $X \times Y$ with marginals μ and ν respectively. Given c, the functional I maps a measure to its cost

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

The Kantorovich problem is to minimize I over $\Pi(\mu, \nu)$.

Lemma 2.1 *Let μ and ν be probability measures defined on Polish spaces X and Y respectively. Then the set of transference plans $\Pi(\mu, \nu)$ is tight.*

Proof. As stated in the introduction, by [2, Theorem 1.3], μ is tight and, analogously, so is ν. Given $\varepsilon > 0$, there exist compact sets $A \subseteq X$ with $\mu(A) > 1 - \varepsilon/2$, and $B \subseteq Y$ with $\nu(B) > 1 - \varepsilon/2$. For any $\pi \in \Pi(\mu, \nu)$ we have

$$\pi((X \times Y) \setminus (A \times B)) \leq \pi((X \setminus A) \times Y) + \pi(X \times (Y \setminus B)) = \mu(X \setminus A) + \nu(Y \setminus B) < \varepsilon.$$

Since $A \times B$ is compact, tightness is proved.

Since $\Pi(\mu, \nu)$ is tight, it is compact by Prokhorov's theorem ([2, Theorem 5.1]). This means that any infinite sequence in $\Pi(\mu, \nu)$ admits partial limits in $\Pi(\mu, \nu)$, providing the basis for the existence proof of an optimal transference plan.

Lemma 2.2 *If X and Y are Polish and c is lower semi-continuous, the Kantorovich problem admits a minimizer.*

Proof. Let $\{\pi_k\}$ be a minimizing sequence for $I(\cdot)$ on $\Pi(\mu, \nu)$, then it admits a subsequence that converges weakly to a probability measure $\pi^* \in \Pi(\mu, \nu)$. We assume without loss of generality that π_k itself converges to π^*. With Proposition 1.4 as justification, write $c = \sup c_l$ where c_l is a monotone sequence of continuous bounded functions. Then the weak convergence together with Portmanteau theorem give

$$\int_{X \times Y} c_l d\pi^* = \lim_{k \to \infty} \int_{X \times Y} c_l d\pi_k \leq \liminf_{k \to \infty} \int_{X \times Y} c d\pi_k, \qquad l \in \mathbb{N}.$$

By definition of the sequence $\{\pi_k\}$ and the monotone convergence theorem,

$$\int_{X\times Y} c d\pi^* = \lim_{l\to\infty} \int_{X\times Y} c_l d\pi^* \leq \liminf_{k\to\infty} \int_{X\times Y} c d\pi_k = \inf_{\tau\in\Pi(\mu,\nu)} \int_{X\times Y} c d\tau.$$

It follows that π^* is indeed a minimizer for $I(\cdot)$.
When c is continuous, $I : P(X \times Y) \to [0, \infty]$ is continuous with respect to weak convergence; this is just the Portmanteau theorem. The last proof shows that when c is lower semi-continuous, so is I.
Define

$$\Phi_c \stackrel{\text{def}}{=} \{(\varphi, \psi) \in L_1 : \varphi(x) + \psi(y) \leq c(x, y) \quad \text{for } \mu\text{-almost any } x \text{ and } \nu\text{-almost any } y\},$$

where we mean of course that $\varphi \in L_1(\mu)$ and $\psi \in L_1(\nu)$. For (φ, ψ) let

$$J(\varphi, \psi) = \int_X \varphi d\mu + \int_Y \psi d\nu. \tag{2.1}$$

The functions φ and ψ are the unknowns of the maximization problem

$$\sup J(\varphi, \psi), \qquad (\varphi, \psi) \in \Phi_c. \tag{2.2}$$

Notice that the value of J is insensitive to measure zero sets: given $(\varphi, \psi) \in \Phi_c$, we can change φ and ψ on μ and ν-measure zero sets so that

$$\varphi(x) + \psi(y) \leq c(x, y) \qquad \forall x, y,$$

without changing the value of $J(\varphi, \psi)$ (by allowing the functions (φ, ψ) to take infinite negative values). The problem (2.2) is the **dual problem** of the Kantorovich problem (1.4). In the context of duality, the original problem is often called **primal problem**.
The relation between these two problems is given by the following fundamental result of the theory of optimal transportation.

Theorem 2.3 (Kantorovich duality) *Let μ and ν be probability measures on the Polish spaces X and Y respectively. Let $c : X \times Y$ be a nonnegative lower semi-continuous function, then*

$$\inf_{\pi\in\Pi(\mu,\nu)} I(\pi) = \sup_{(\varphi,\psi)} J(\varphi, \psi),$$

where $I(\pi)$ is defined by (1.4) and $J(\varphi, \psi)$ by (2.1).

Proof. The proof that the infimum is at least as large as the supremum is easy. Invoking the same idea in the proof of Lemma 2.1, we see that a property that holds for μ-almost any x and ν-almost any y also holds π-almost surely if $\pi \in \Pi(\mu, \nu)$. Therefore, if $(\varphi, \psi) \in \Phi_c$ and $\pi \in \Pi(\mu, \nu)$, $\varphi(x) + \psi(y) \leq c(x, y)$ π-almost surely and by Proposition 1.1

$$J(\varphi, \psi) = \int_{X\times Y} (\varphi(x) + \psi(y)) d\pi(x, y) \leq \int_{X\times Y} c(x, y) d\pi(x, y) = I(\pi).$$

The proof of the reverse inequality uses a minimax argument: this is an argument of the form

$$\inf_{a\in A}\sup_{b\in B} f(a,b) = \sup_{b\in B}\inf_{a\in A} f(a,b). \tag{2.3}$$

The equality (2.3) does not always hold: if $A = B = \{1,2\}$ and $f(a,b) = \delta_{ab} = 1\{a = b\}$ then the left hand side equals 1 while the right side equals 0. In any case "$\geq$" holds, while the "$\leq$" requires some conditions on the function f. We give a proof for the Kantorovich duality that uses a minimax argument and sketch its justification later. The idea is to "regularize" both problems by considering all the positive measures on $X\times Y$, $M_+(X\times Y)$, and all functions $\varphi \in L_1(\mu)$, $\psi \in L_1(\nu)$ while introducing penalties if $(\varphi,\psi) \notin \Phi_c$ and $\pi \notin \Pi(\mu,\nu)$.

It is clear that

$$I \stackrel{\text{def}}{=} \inf_{\pi\in\Pi(\mu,\nu)} I(\pi) = \inf_{\pi\in M_+(X\times Y)} \begin{cases} I(\pi) & \pi \in \Pi(\mu,\nu) \\ \infty & \text{otherwise,} \end{cases}$$

and, by Proposition 1.1,

$$\sup_{\varphi\in L_1(\mu),\psi\in L_1(\nu)} J(\varphi,\psi) - \int_{X\times Y} (\varphi(x)+\psi(y))d\pi(x,y) = \begin{cases} 0 & \pi \in \Pi(\mu,\nu) \\ \infty & \text{otherwise.} \end{cases}$$

(If $\pi \notin \Pi(\mu,\nu)$ then $(\varphi,\psi) \in L_1$ exist so that the expression inside the supremum operator is nonzero; multiplying both functions by an arbitrary constant gives the result.) Combining the two equalities, we have

$$I = \inf_{\pi\in M_+(X\times Y)} \left(I(\pi) + \sup_{\varphi\in L_1(\mu),\psi\in L_1(\nu)} \left(J(\varphi,\psi) - \int_{X\times Y} (\varphi(x)+\psi(y))d\pi(x,y) \right) \right).$$

Invoking a minimax principle and replacing $I(\pi)$ by its definition, the right side equals

$$\sup_{(\varphi,\psi)\in L_1} \left(J(\varphi,\psi) + \inf_{\pi\in M_+(\mu,\nu)} \left(\int_{X\times Y} (c(x,y) - \varphi(x) - \psi(y))d\pi(x,y) \right) \right).$$

Specifying $\pi \equiv 0$ shows that the infimum can never take a positive value; whenever $(\varphi,\psi) \in \Phi_c$ neither can it take a negative one. If $c(x,y) < \varphi(x) + \psi(y)$ for some $x \in X$, $y \in Y$, the family of positive measures $\{\lambda\delta_{(x,y)}\}_{\lambda>0}$ yields an infinite negative value for the infimum. Therefore we must consider only functions satisfying $\varphi(x)+\psi(y) \leq c(x,y)$ and, as J is insensitive to the values of the functions on measure zero, the above expression reduces to

$$\sup_{(\varphi,\psi)\in\Phi_c} J(\varphi,\psi),$$

which is what was to be proved.

Before justifying the minimax argument observe that the dual problem (2.2) can be reduced to finding only one function, and not two. Given $\varphi \in L_1(\mu)$, ψ should be the

maximal function satisfying $\varphi(x) + \psi(y) \leq c(x,y)$. Clearly, the maximal ψ under this constraint is

$$\varphi^c(y) = \inf_{x \in X} c(x,y) - \varphi(x). \tag{2.4}$$

Functions of the form (2.4) are said to be **c-concave**, and play an important role in the Kantorovich duality theory. A-priori, it is not clear that φ^c is measurable; note however that when c is uniformly continuous, so is φ^c. Of course, we can interchange the roles of ψ and φ, and replace φ by

$$\varphi^{cc}(x) = \inf_{y \in Y} c(x,y) - \varphi^c(y) = \inf_{y \in Y} \sup_{z \in X} c(x,y) - c(z,y) + \varphi(z).$$

$\varphi^{cc}(x) \geq \varphi(x)$ for any $x \in X$ and we have

$$\varphi^{cc}(x) + \varphi^c(y) \leq c(x,y) \qquad \forall x, y.$$

There is no point in continuing further; it is easy to see that $\varphi^{ccc} = \varphi^c$. The pair $(\varphi^{cc}, \varphi^c)$ is a pair of **c-conjugate** functions. This conjugation procedure allows us to increase the value of J without violating the constraints.

We now state without proof the argument proving the minimax argument. Recall that for a topological vector space X, its (topological) **dual** space X^* is the collection of continuous linear functional from X to $\mathbb{R}$. Before stating the result, we need one last definition. Notice the similarity to the definition of φ^c.

Definition 2.4 *Let E be a normed vector space, E^* its dual and $\Theta : E \to \mathbb{R} \cup \{\infty\}$ be convex. The **Legendre–Fenchel transform** of Θ is a function $\Theta^* : E^* \to \mathbb{R} \cup \{-\infty\}$ defined by*

$$\Theta^*(z^*) = \sup_{z \in E} (z^*(z) - \Theta(z)).$$

The definition implies $\Theta(z) + \Theta^*(z^*) \geq z^*(z)$ for any $z \in E$, $z^* \in E^*$.

Theorem 2.5 (Fenchel–Rockafellar duality) *Let E be a normed vector space, E^* its dual and $\Theta, \Xi : E \to \mathbb{R} \cup \{\infty\}$ two convex functions with Legendre–Fenchel transforms Θ^*, Ξ^*. Assume that there exists $z_0 \in E$ such that Θ is continuous at z_0 and $\Theta(z_0) + \Psi(z_0)$ is finite. Then*

$$\inf_{z \in E} (\Theta(z) + \Xi(z)) = \sup_{z^* \in E^*} \left(-\Theta^*(-z^*) - \Xi^*(z^*)\right), \tag{2.5}$$

and the right hand side is a maximum.

Theorem 2.5 is a minimax theorem, because the right hand side of (2.5) equals

$$\sup_{z^* \in E^*} \inf_{z, z' \in E} \left(z^*(z - z') + \Theta(z) + \Xi(z')\right),$$

and reversing the order of the supremum and the infimum operators yields the left hand side of (2.5). A proof of the theorem is given in [14, p. 24].

With Theorem 2.5 at hand, we can prove the Kantorovich duality when X and Y are compact and c is continuous; the more general case will follow by approximation arguments. $X \times Y$ is also compact and Polish, and we consider the normed vector space

$$E = C_b(X \times Y), \qquad \|f\| = \sup_{x,y} |f(x,y)|, \qquad f \in E.$$

By the Riesz representation theorem and the Polishness of $X \times Y$, the dual space E^* is equivalent to $M(X \times Y)$, the set of finite signed Borel measures on $X \times Y$. Define $\Theta, \Xi : E \to \mathbb{R} \cup \{\infty\}$ by

$$\Theta(u) \stackrel{\text{def}}{=} \begin{cases} 0 & u(x,y) \geq -c(x,y) \quad \forall x, y \\ \infty & \text{otherwise,} \end{cases}$$

and

$$\Xi(u) \stackrel{\text{def}}{=} \begin{cases} J(\varphi, \psi) & u(x,y) = \varphi(x) + \psi(y) \quad \forall x, y \\ \infty & \text{otherwise.} \end{cases}$$

It is easy to see that Θ and Ξ are convex, and that the hypotheses of Theorem 2.5 are satisfied when the function z_0 is constant 1. Therefore (2.5) holds and, interpreting the elements of E^* as measures, a little bit of (easy) algebra recovers the Kantorovich duality. Theorem 2.5 guarantees that the infimum of I on $\Pi(\mu, \nu)$ is attained, but we already knew that by Lemma 2.2.

The next step is to relax the compactness assumption of X and Y. We know already that a minimizer π^* exists and, using tightness, we can find compact sets $A \subseteq X$ and $B \subseteq Y$ with $\mu(A) > 1 - \varepsilon$ and $\nu(B) > 1 - \varepsilon$ and consequently $\pi^*(A \times B) > 1 - 2\varepsilon$. Restricting π^* to $A \times B$ and multiplying by the factor $1/\pi^*(A \times B)$, we obtain a probability measure on the compact space $A \times B$, with marginal distributions μ_0 and ν_0. By the Kantorovich duality for $A \times B$, we have

$$\inf_{\pi \in \Pi(\mu_0, \nu_0)} I(\pi) = \sup_{(\varphi,\psi) \in \Phi_c} J_0(\varphi, \psi) \stackrel{\text{def}}{=} \int_X \varphi d\mu_0 + \int_Y \psi d\nu_0. \tag{2.6}$$

It quite easy to see that the left hand side is infinitesimally close to $\inf I$ over $\Pi(\mu, \nu)$. The more complex approximation argument is to show that the right hand side is close to $\sup J(\varphi, \psi)$. The details are given in [14, p. 28–31].

The last part of the proof consists of relaxing the restriction on the cost function c, assuming only lower semi-continuity. To this end, we write $c = \sup c_n$ where $c_n \nearrow c$ are nonnegative, bounded and uniformly continuous functions. The approximation is now rather simple: for each n, let π_n be the minimizer of $I_n(\pi) \stackrel{\text{def}}{=} \int c_n d\pi$ over $\Pi(\mu, \nu)$. Then by the Kantorovich duality for c_n,

$$I_n(\pi_n) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c_n d\pi = \sup_{(\varphi,\psi) \in \Phi_{c_n}} J(\varphi, \psi) \stackrel{\text{def}}{=} J_n.$$

Our goal is to show that

$$J \stackrel{\text{def}}{=} \sup_{(\varphi,\psi)\in\Phi_c} J(\varphi,\psi) \geq \inf_{\pi\in\Pi(\mu,\nu)} I(\pi), \qquad I(\pi) \stackrel{\text{def}}{=} \int_{X\times Y} c d\pi,$$

since the reverse inequality has already been established. Since $\Phi_{c_n} \subseteq \Phi_{c_{n+1}} \subseteq \Phi_c$, J_n is a monotone sequence with $\lim J_n \leq J$. By tightness of $\Pi(\mu,\nu)$, up to the extraction of a subsequence, $\pi_m \xrightarrow{\text{D}} \pi^*$, and by the monotone convergence theorem $I_n(\pi) \nearrow I(\pi)$ for any $\pi \in M_+(X\times Y)$. Thus

$$I \leq I(\pi^*) = \lim_{n\to\infty} I_n(\pi^*) = \lim_{n\to\infty}\lim_{m\to\infty} I_n(\pi_m) \leq \lim_{n\to\infty}\lim_{m\to\infty} I_m(\pi_m) = \lim_{m\to\infty} J_m,$$

or

$$\inf_{\pi\in\Pi(\mu,\nu)} I(\pi) \leq I(\pi^*) \leq \lim_{n\to\infty} J_n \leq J,$$

completing the proof. Note also that it follows that π^* is a minimizer of I. The proof of the Kantorovich duality is now complete.

We have shown, either by the Fenchel–Rockafellar duality or by tightness argument, that the infimum of I is in fact a minimum. A few words should be said about the supremum of J. Notice that it was never assumed that the infimum of I is finite—it is certainly possible that any transference plan incurs an infinite cost, and the Kantorovich duality still holds for these cases too. Since we require the functions φ and ψ to be in L_1, it is clear that $J(\varphi,\psi)$ is always finite and the supremum of J is not a maximum. It can be shown (see e.g. [1]) that if c is bounded by L_1 functions

$$0 \leq c(x,y) \leq c_X(x) + c_Y(y), \qquad c_X \in L_1(\mu), \quad c_Y \in L_1(\nu), \tag{2.7}$$

then the supremum of J is a maximum, and is attained by c-conjugate functions. The argument of the proof involves the notion of c-monotonicity, which we introduce in Section 7. Note that the condition (2.7) implies finiteness of $I(\pi)$ for any coupling $\pi \in \Pi(\mu,\nu)$. Uniqueness in the dual problem cannot hold, because $J(\varphi - s, \psi + s) = J(\varphi,\psi)$ for any $s \in \mathbb{R}$.

2.2 Examples of dual problems

In this subsection, we recall the examples from the introduction and present their dual problems.

For the trivial example 0 where μ is a dirac measure defined on $X = \{x\}$, the cost function $c(x,y)$ is simply $c(y)$, since x is unique. Specifying $\varphi = 0$ and $\psi = c$ yields

$$J(\varphi,\psi) = \int_Y c(x,y)dy = \int_{X\times Y} c(x,y)d\pi, \qquad \pi = \mu\otimes\nu.$$

This proves optimality of φ and ψ.

Example 1: Let λ be Lebesgue measure on $\mathbb{R}$, $X = Y = [0, n+1]$, $c(x,y) = |y-x|^p$ $(p > 0)$, $\mu = \frac{1}{n}\lambda|_{[0,n]}$ and $\nu = \frac{1}{n}\lambda|_{[1,n+1]}$. Consider the transport maps

$$T_1(x) = x + 1 \qquad \text{and} \qquad T_2(x) = \begin{cases} x & 1 < x \le n \\ x + n & 0 \le x \le 1 \end{cases}$$

with costs 1 and n^{p-1} respectively. The dual problem is to find $\varphi : [0,n] \to \mathbb{R}$ and $\psi : [1, n+1] \to \mathbb{R}$ such that $\varphi(x) + \psi(y) \le |y-x|^p$ and $J(\varphi, \psi)$ is maximal. For $p > 1$ let $\varphi(x) = -px$ for $x \in [0,n]$. Its c-conjugate is

$$\psi(y) = \varphi^c(y) = \inf_{0 \le x \le n} |y-x|^p + px = 1 - p + py, \qquad y \in [1, n+1],$$

since the infimum is attained at $x = y - 1 \in [0,n]$. Then $\varphi(x) + \psi(y) \le |y-x|^p$, and

$$J(\varphi,\psi) = \frac{1}{n}\int_1^{n+1} (1-p+py)dy - \frac{1}{n}\int_0^n pxdx = (1-p) + \frac{p}{n}\frac{(n+1)^2 - 1 - n^2}{2} = 1.$$

Consequently, T_1 is optimal for the primal, and the pair (φ, ψ) is optimal for the dual.
When $p < 1$ the situation is more complex. Since the cost function is a metric, all the shared mass can stay in place and one only needs to optimally transport $[0,1]$ to $[n, n+1]$ and divide the cost by n (see the Kantorovich–Rubinstein theorem in the next subsection). Letting $\varphi(x) = (n+1-2x)^p/2$ for $x \in [0,1]$ we see that

$$\psi(y) = \varphi^c(y) = \inf_{x \in [0,1]} (y-x)^p - (n+1-2x)^p/2 = (2y-n-1)^p/2, \qquad y \in [n, n+1]$$

since the infimum is attained at $x = n + 1 - y \in [0,1]$. The functional J equals

$$J(\varphi,\psi) = \frac{1}{2n}\int_n^{n+1} (2y-n-1)^p dy + \frac{1}{2n}\int_0^1 (n+1-2x)^p dx = \frac{(n+1)^{p+1} - (n-1)^{p-1}}{2n(p+1)},$$

which is precisely the cost of the map $x \mapsto n+1-x$:

$$\frac{1}{n}\int_0^1 (n+1-2x)^p dx = \frac{(n+1)^{p+1} - (n-1)^{p-1}}{2n(p+1)}.$$

It follows that the map

$$T_3(x) = \begin{cases} x & 1 < x \le n \\ n+1-x & 0 \le x \le 1 \end{cases},$$

optimally transports μ to ν. Observe that without the reduction to disjoint segments the functions φ and ψ are not even well-defined!
See Subsection 5.2 to see how to find the functions (φ, ψ) and their connection to the optimal map.
When $p = 1$, both constructions work equally well, because the resulting functions differ by a constant. Note that when the segments are disjoint, the cost function is already of

the form $y - x = \varphi(x) + \psi(y)$ and the dual problem is trivial (though not as degenerate as the primal!).

Example 2: $\mu = (1/3)\delta\{x_1\} + (2/3)\delta\{x_2\}$, $\nu = (1/3)\delta\{y_1\} + (2/3)\delta\{y_2\}$, $c_{ij} = c(x_i, y_j)$. Transference plans consist of transporting a_{11} mass from x_1 to y_1, $a_{11} + 1/3$ from x_2 to y_2, and $(1/3 - a_{11})$ from x_1 to y_2 and from x_2 to y_1, and the operator $I(\pi)$ is linear with $a_{11} \in [0, 1/3]$. The dual problem is to find four real numbers φ_1, φ_2, ψ_1 and ψ_2 such that $\varphi_i + \psi_j \leq c_{ij}$ and $(\varphi_1 + 2\varphi_2 + \psi_1 + 2\psi_2)/3$ is maximal. A moment's reflection suffices to conclude that $\varphi_2 + \psi_2 \leq c_{22}$ should hold as equality; choose any such combination of φ_2 and ψ_2, and $\psi_1 = c_{21} - \varphi_2$. Thus the inequalities concerning φ_2 are satisfied as equalities. Now choose $\varphi_1 = \min(c_{11} - \psi_1, c_{12} - \psi_2)$. Easy algebra show that this minimum is determined by the sign of $c_{11} + c_{12} - c_{21} - c_{12}$ and that in both cases $J(\varphi, \psi) = I(a_{11})$ for the optimal a_{11}.

Example 3: $\mu = (1/n)\sum_{i=1}^{n} \delta\{x_i\}$ and $\nu = (1/n)\sum_{j=1}^{n} \delta\{y_j\}$. The dual problem is to find $2n$ real numbers $\varphi_1, \ldots, \varphi_n$ and $\psi_1, \ldots, \psi_n$ such that

$$\forall i, j \qquad \varphi_i + \psi_j \leq c_{ij}, \tag{2.8}$$

and maximizing

$$J(\varphi, \psi) = \frac{1}{n}\sum_{i=1}^{n} \varphi_i + \frac{1}{n}\sum_{j=1}^{n} \psi_j.$$

Let $[n] = \{1, \ldots, n\}$. If we can find a permutation $\sigma : [n] \to [n]$ such that (2.8) is satisfied with equality when $j = \sigma(i)$, then

$$J(\varphi^\sigma, \psi^\sigma) = \frac{1}{n}\sum_{i=1}^{n} \varphi_i^\sigma + \psi_{\sigma(i)}^\sigma = \frac{1}{n}\sum_{i=1}^{n} c_{i\sigma(i)} = I(\sigma),$$

where the right hand side is the cost incurred from the transport map σ. It follows from Theorem 2.3 that σ is optimal for the Monge–Kantorovich problem and the pair $(\varphi^\sigma, \psi^\sigma)$ is optimal for the dual problem. The Hungarian method presented in Section 4 is in fact an algorithm for solving the dual problem, by finding such a permutation.

We omit the dual problem for the fourth and fifth examples, the former for brevity and the latter as it is a private case of Example 1 when $n = 1$.

2.3 The Kantorovich–Rubinstein theorem

We conclude this section by presenting a special case of Theorem 2.3, when the cost function c is a metric. An immediate corollary will be that in this case, all shared mass can stay in place.

Theorem 2.6 (Kantorovich–Rubinstein) *Let (X, d) be a Polish space and $\mu, \nu \in P(X)$. Let $c : X^2 \to [0, \infty)$ be a lower semi-continuous function satisfying the axioms of a metric. Then*

$$\inf_{\pi \in \Pi(\mu,\nu)} I(\pi) = \sup_{\|\varphi\|_{Lip} \leq 1} \int_X \varphi d(\mu - \nu),$$

where

$$\|\varphi\|_{Lip} \stackrel{\text{def}}{=} \sup_{x \neq y} \frac{|\varphi(x) - \varphi(y)|}{d(x,y)}$$

is the smallest K such that φ is K-Lipschitz.

Remark 1 *Note that except its lower semi-continuity, c has nothing to do with the metric of the space d. For example, convergence with respect to c is neither stronger nor weaker than convergence with respect to d.*

Proof. We only prove the theorem for bounded c. If c is unbounded take $c_n = g_n \circ c$ for $g_n(z) = nz/(n+z)$, then c_n is a bounded lower semi-continuous metric and the theorem applies to c_n. The same argument in the last step of the proof of the Kantorovich duality proves the result for c.
By the Kantorovich duality, what we have to show is that

$$\sup_{(\varphi,\psi)\in\Phi_c} J(\varphi,\psi) = \sup_{\|\varphi\|_{Lip}\leq 1} J(\varphi,-\varphi).$$

When c is bounded, any 1-Lipschitz function φ is bounded and therefore integrable, and furthermore $(\varphi,-\varphi) \in \Phi_c$. Therefore the right hand side cannot exceed the left hand side. For the other inequality, observe that the left hand side equals

$$\sup_{\varphi \text{ bounded}} J(\varphi^{cc},\varphi^c), \qquad \varphi^c(y) = \inf_{x\in X} c(x,y) - \varphi(x), \quad \varphi^{cc}(x) = \inf_{y\in Y} c(x,y) - \varphi^c(y).$$

When φ is bounded, φ^c is well defined, and when comparing $\varphi^c(y)$ with $\varphi^c(z)$ we see that the expression inside the infimum operator differs by $|c(x,y) - c(x,z)| \leq c(y,z)$ for any x; from this it follows that φ^c is 1-Lipschitz. As a general fact, if ψ is 1-Lipschitz then $\psi^c \geq -\psi$, and in any case $\psi^c(x) \geq c(x,x) - \psi(x)$. For our c it follows that $\psi^c = -\psi$ and as φ^c is 1-Lipschitz, $\varphi^{cc} = -\varphi^c$. Thus

$$\sup_{\varphi \text{ bounded}} J(\varphi^{cc},\varphi^c) = \sup_{\varphi \text{ bounded}} J(-\varphi^c,\varphi^c) \leq \sup_{\|\varphi\|_{Lip}\leq 1} J(\varphi,-\varphi),$$

and the result follows.

It is clear that the Kantorovich duality and the Kantorovich–Rubinstein theorem can be extended to finite measures with $\mu(X) = \nu(Y)$, since $I(\alpha\pi) = \alpha I(\pi)$, $\alpha \geq 0$. Therefore the following corollary makes sense. It it immediate from the right hand side of the Kantorovich–Rubinstein theorem, because we can also impose φ to be bounded.

Corollary 2.7 *Let X be Polish and $\mu,\nu,\sigma \in M_+(X)$ such that $\mu(X) = \nu(X)$. Let c be a metric that is lower semi-continuous with respect to the metric of X, then*

$$\inf_{\tilde{\pi}\in\Pi(\mu+\sigma,\nu+\sigma)} I(\tilde{\pi}) = \inf_{\pi\in\Pi(\mu,\nu)} I(\pi).$$

This means that there exists an optimal transference plan that leaves all common mass in place.

In general, when $c(x,x) \equiv 0$, the "$\leq$" in Corollary 2.7 is easy: given $\pi \in \Pi(\mu,\nu)$, we can construct $\tilde{\pi} \in \Pi(\mu+\sigma, \nu+\sigma)$ by coupling σ with itself using the identity map

$$\tilde{\pi}(A \times B) = \pi(A \times B) + \sigma(A \cap B).$$

Then $I(\tilde{\pi}) = I(\pi)$. Obviously, $\Pi(\mu+\sigma, \nu+\sigma)$ consists of elements that were not constructed this way, so only inequality is proven. Equality may fail if c is not a metric. Indeed, the cost function $c(x,y) = |x-y|^p$, $p > 1$ is not a metric. When $X = \mathbb{R}$, $\mu = \delta_0$, $\nu = \delta_2$ and $\sigma = \delta_1$, the left hand side equals 2 while the right hand side is $2^p > 2$. The optimal coupling for $\Pi(\mu+\sigma, \nu+\sigma)$ does *not* couple σ with itself.

When c is a metric, so is c^p for $0 < p < 1$, by the following elementary lemma. A metric $d : X^2 \to \mathbb{R}$ satisfies the **strict triangle inequality** if $d(x,y) + d(y,z) \geq d(x,z)$ for any $x, y, z \in X$, and strict inequality holds if x, y and z are distinct.

Lemma 2.8 *Let d be a metric and $h : [0,\infty) \to [0,\infty)$ be concave where $h(x) = 0$ if and only if $x = 0$. Then $h \circ d$ is a metric. If h is strictly concave then $h \circ d$ satisfies the strict triangle inequality.*

Proof. Concavity with $h(0) = 0$ imply that $h(a) \geq \frac{a}{a+b} h(a+b)$. Interchanging a and b and summing up we have $h(a+b) \leq h(a) + h(b)$, so h is **subadditive**. In addition such h has to be nondecreasing. Thus

$$h(d(x,y)) + h(d(y,z)) \geq h(d(x,y) + d(y,z)) \geq h(d(y,z)).$$

If h is strictly concave then $h(a+b) < h(a) + h(b)$ unless a or b vanish, and h is strictly increasing. Then the left inequality above is strict when x, y and z are distinct.
If $0 < p < 1$, then $|x-y|^p$ satisfies the strict triangle inequality, because $x \mapsto x^p$ is strictly concave. Gangbo and McCann [6] show that for such cost functions, all the shared mass *has* to stay in place, a stronger result than Corollary 2.7.

3 Linear programming

3.1 Relation to the Monge–Kantorovich problem

As stated in the Example 3 in the introduction, solving the Monge problem for two discrete measures $\mu = (1/n)\sum_{i=1}^{n} \delta_{p_i}$ and $\nu = (1/n)\sum_{j=1}^{n} \delta_{q_j}$ reduces to finding a permutation σ that minimizes

$$c(\sigma) = \frac{1}{n}\sum_{i=1}^{n} c\left(p_i, q_{\sigma(i)}\right).$$

The Kantorovich problem is to find a bistochastic matrix x that minimizes

$$c(x) = \frac{1}{n}\sum_{i,j=1}^{n} x_{ij} c(p_i, q_j) = \sum_{i,j=1}^{n} x_{ij} c_{ij}, \qquad c_{ij} = c\left(p_i - q_j\right),$$

Ignoring the factor $1/n$ and considering c and x as n^2-dimensional vectors, one can express the Kantorovich problem as finding

$$\min_{x \in \mathbb{R}^{n^2}} c^t x,$$

where the vector c is given by $c_{ij} = c(p_i, q_j)$, and the vector x should have nonnegative entries and satisfy

$$\forall i, j \quad \sum_{i=1}^{n} x_{ij} = \sum_{j=1}^{n} x_{ij} = 1.$$

The constraints are linear in x and can be written in matrix form. If $n = 3$ for example, the problem is

$$\min c_{11}x_{11} + c_{12}x_{12} + c_{13}x_{13} + c_{21}x_{21} + c_{22}x_{22} + c_{23}x_{23} + c_{31}x_{31} + c_{32}x_{32} + c_{33}x_{33}$$

subject to $x_{ij} \geq 0$ and

$$\begin{pmatrix} 1 & 1 & 1 & & & & & & \\ & & & 1 & 1 & 1 & & & \\ & & & & & & 1 & 1 & 1 \\ 1 & & & 1 & & & 1 & & \\ & 1 & & & 1 & & & 1 & \\ & & 1 & & & 1 & & & 1 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{31} \\ x_{32} \\ x_{33} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

For general n, the constrains can be expressed by the vector equation $Ax = b$, where
(1) $x \in \mathbb{R}_{+}^{n^2}$;
(2) the matrix $A \in M_{2n,n^2}(\mathbb{R})$ has n^2 columns and $2n$ rows; and
(3) $b = (1, \ldots, 1) \in \mathbb{R}^{2n}$.

The matrix A is of the form

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

where the lower part consists of n copies of the identity matrix in M_n.

Remark 2 *The rank of A is $2n-1$ for any $n \geq 1$.*

Proof. Denote the rows of A by $A^1, \dots, A^{2n}$, then the $\mathbb{R}^{n^2}$-equality $\sum_{i=1}^{n} A^i = \mathbf{1}_{n^2} = \sum_{i=n+1}^{2n} A^i$ implies that the rank of A cannot exceed $2n-1$. It is however easy to see that the first $2n-1$ rows are linearly independent: if $\sum_{i=1}^{2n-1} \lambda_i A^i = 0$ then $\lambda_1 = 0$ because in the n-th column only the first element is non zero. $\lambda_2 = 0$ as well because in the $2n$-th column only the second element is non zero. Similarly, $\lambda_3 = \cdots = \lambda_n = 0$. The result now follows, since the bottom n rows of A are linearly independent. One can similarly show that *any* collection of $2n-1$ rows is linearly independent, either directly or by using the symmetry between the rows of A (think about permuting the vector x).
As we shall see later, the theory of linear programming often assumes that the constraints matrix is of full row rank. By Remark 2, we can obtain such a matrix from A by removing one of its rows.
Having seen the relationship between the discrete Monge–Kantorovich problem and linear programming, the rest of this section surveys the basic ideas and methods of the latter. In Section 4 we present a method that takes advantage of the special structure of the matrix A in order to solve the problem efficiently.

3.2 Introduction to linear programming

In this subsection and those that follow it, we develop the basics of the theory of linear programming. The presentation here is self-contained (except when dealing with degenerate basic solutions), but brief; the reader is referred to e.g. [9] for more examples and a more extensive discussion.
An **unconstrained optimization problem** arises when one wishes to maximize or minimize a real-valued function, f, say, over the Euclidean space $\mathbb{R}^d$. f is called the **objective function** of the problem. In consumer theory in economics, for instance, f can be a utility function and $x \in \mathbb{R}^d$ may represent some commodities or products. $f(x)$ is then the utility the consumer obtains from the combination of products x. The optimization problem the consumer wishes to solve (often referred to in the literature as *the consumer problem*) is finding the combination of commodities x^* where the utility function attains its maximal value: find $x^* \in \mathbb{R}^d$ such that $f(x^*) \geq f(x)$ for any $x \in \mathbb{R}^d$.

Another example comes, too, from economics. Imagine that f represents the net profit of a firm from producing certain types of products. If products are manufactured beyond the market demand and end up not being sold, the net profit can be increased by producing less. Similarly, if demand exceeds market supply, it might be advantageous for the firm to increase production. The firm seeks to find production quantities $x^* \in \mathbb{R}^d$ which maximize the profit function f.

In some cases the choice of the variable x is restricted; some **constraints** are present. In this case one talks about **constrained problems**. For instance, the consumer who wishes to maximize his utility cannot buy huge quantities from the products because his budget might be limited. If a_i is the price of one unit of product i and the consumer can only spend B, the maximization has to be done under the constraint that $a_1x_1+a_2x_2+\cdots+a_dx_d \leq B$. The firm in the example above may have limited storage space and (assuming that all products have the same volume) then a constraint of the form $x_1 + x_2 + \cdots + x_d \leq S$ appears. We will restrict attention to constraints of the form $g(x) \leq \alpha$ or $h(x) = \beta$. Clearly, several constraints may be imposed simultaneously.
A classical example of a geometrically constrained problem is: given a real number $R > 0$, find a rectangle with perimeter R and maximal area. This optimization problem can be formulated as

$$\max ab \qquad \text{subject to} \quad 2a + 2b = R \quad \text{and} \quad a, b \geq 0.$$

Here a and b are the sides of the rectangle, $f(a, b) = ab$ is its area and $2a+2b = h(a, b) = R$ is the perimeter constraint.

Finally, we can present the particular case in which we are interested in this section. When we write $x \geq y$ for $x, y \in \mathbb{R}^d$, we mean that $x_i \geq y_i$ for $i = 1, \ldots, d$.

Definition 3.1 *A **linear program** is a constrained optimization problem in which both the objective function and the constraints are linear in the unknowns. It is given in **standard form** if it is formulated as*

$$\min c^t x \quad \textit{subject to} \quad Ax = b \quad \textit{and} \quad x \geq 0,$$

*where $c \in \mathbb{R}^n$, $A \in M_{m,n}$ and $b \in \mathbb{R}^m_+$ are given. $x \in \mathbb{R}^n$ are the variables of the problem which are to be determined. The function $x \mapsto c^t x$ is the **objective function** of the program. $x^* \in \mathbb{R}^n$ is a **solution** if it satisfies $Ax^* = b$, and is **feasible** if in addition $x^* \geq 0$. It is **optimal** if $c^t x^* = \min\{c^t x : x \in \mathbb{R}^n, Ax = b, x \geq 0\}$.*

$Ax = b$ is a compact way of writing the m simultaneous linear constraints

$$A^1x = b_1 \quad A^2x = b_2 \quad \ldots \quad A^mx = b_m,$$

where $A^i \in \mathbb{R}^n$ is the i-th row of A and $b_i \in \mathbb{R}_+$.

Any linear program can be transformed into the standard form. Firstly, maximizing $c^t x$ is equivalent to minimizing $(-c)^t x$.
An inequality constraint of the form $a^t x \leq \beta$ is equivalent to $\tilde{a}^t \tilde{x} = \beta$ where $\tilde{a} = (a, 1)$, and $\tilde{x} = (x, y)$ for a scalar variable $y \geq 0$. The vector c has to be extended to $\tilde{c} = (c, 0)$ and then $\tilde{c}^t \tilde{x} = c^t x$. The inequality is thus transformed into equality, with the expense of adding an extra variable. This variable is called *slack variable*.
A constraint of the form $a^t x \geq \beta$ can be similarly transformed into an equality constraint by $\tilde{a} = (a, -1)$ and $\tilde{x} = (x, y)$, $y \geq 0$. The extra variable y is then called *surplus variable*.
If some coordinate of the vector b is negative, one can negate the corresponding row of A: $a^t x = \beta$ is equivalent to $(-a)^t x = -\beta$.
In some programs the variables are not required to be nonnegative; the constraint $x_1 \geq 0$, for instance, might be missing. There are several ways to transform such programs to the standard form. One option is to replace x_1 by $x_1 = u - v$, where u and v are nonnegative real numbers. This adds many solutions, however, since the choice of u and v is not unique.
Another method is to use one of the constraints $a^t x = \beta$ in order to express x_1 as a function of the other variables. This has the advantage of eliminating one variable and one constraint from the program (if this is impossible, x_1 appears nowhere in the constraints. If $c_1 = 0$ then x_1 appears nowhere in the problem and can be eliminated; if $c_1 \neq 0$, no optimal solution exists because we can choose x_1 such that $x_1 c_1$ as well as the objective function take arbitrarily large negative values).

Using some manipulations, one can even deal with some nonlinear objective functions. As an example, consider the problem

$$\begin{array}{c} \min |x| + |y| + |z| \\ \text{subject to} \quad x + y \leq 1 \quad \text{and} \quad 2x + z = 3. \end{array} \tag{3.1}$$

The objective function of (3.1) is nonlinear, but it is linear on the eight subsets of $\mathbb{R}^3$ that correspond to the signs taken by each of the three variables. One way to solve (3.1) by linear programming is to split it into the eight linear programs

$$\begin{array}{c} \min x + y + z \\ \text{subject to} \quad x + y \leq 1 \qquad 2x + z = 3 \qquad x, y, z \geq 0 \end{array}$$

$$\begin{array}{c} \min x + y - z \\ \text{subject to} \quad x + y \leq 1 \qquad 2x + z = 3 \qquad x, y \geq 0 \quad z \leq 0 \end{array}$$

$$\begin{array}{c} \min x - y + z \\ \text{subject to} \quad x + y \leq 1 \qquad 2x + z = 3 \qquad x, z \geq 0 \quad y \leq 0 \end{array}$$

etc. and then compare the solutions (in fact only six programs need to be considered, since either x or z has to be positive).
We present an alternative way for recasting (3.1) as a linear program. We begin by dealing with the nonlinearity of the objective function. Introduce new variables u, v and w, that

are to replace $|x|$, $|y|$ and $|z|$ using some constraints. If $u = |x|$, then both $u - x \geq 0$ and $u + x \geq 0$; the linear constraints $u - x \geq 0$ and $u + x \geq 0$ are together equivalent to $u \geq |x|$. Similarly, we add the constraints $v - y \geq 0$, $v + y \geq 0$, $w - z \geq 0$ and $w + z \geq 0$. The resulting optimization problem

$$\begin{array}{cccccc} & & \min u + v + w & & \text{subject to} & \\ x + y \leq 1 & 2x + z = 3 & u \pm x \geq 0 & v \pm y \geq 0 & w \pm z \geq 0 & \end{array} \tag{3.2}$$

is a linear program. Since we wish to minimize u and the only constraint on it is that $u \geq |x|$, any optimal solution must satisfy $u = |x|$. Similarly, any optimal solution satisfies $v = |y|$ and $w = |z|$. Therefore, we are indeed minimizing $|x| + |y| + |z|$ after all: the optimal values of x, y and z corresponding to (3.1) and (3.2) coincide. The feasible values of x, y, and z in the two programs are the same as well. To each feasible combination of x, y and z correspond infinitely many feasible values of u, v and w in (3.2) though only $u = |x|$, $v = |y|$, and $w = |z|$ can possibly be optimal. The nonlinear absolute value operations have been successfully eliminated, at the expense of adding three variables and six constraints to the program.

The program (3.2) can be transformed to its standard form as discussed above: firstly, as z is free we can eliminate it from the program using the constraint $z = 3 - 2x$. Having no other equality constraints, we cannot apply this technique to eliminate x and y; we resort to writing $x = x_1 - x_2$, $y = y_1 - y_2$ for nonnegative x_1, x_2, y_1 and y_2:

$$\begin{array}{cccc} & \min u + v + w & \text{subject to} & \\ x_1 - x_2 + y_1 - y_2 \leq 1 & u - x_1 + x_2 \geq 0 & v - y_1 + y_2 \geq 0 & w + 2x_1 - 2x_2 \geq 3 \\ & u + x_1 - x_2 \geq 0 & v + y_1 - y_2 \geq 0 & w - 2x_1 + 2x_2 \geq -3 \end{array}$$

and $x_1, x_2, y_1, y_2 \geq 0$.

Next, we multiply the last constraint by -1 so that the scalar in the right hand side becomes nonnegative: $2x_1 - 2x_2 - w \leq 3$. To transform the inequalities into equalities, we introduce five surplus variables $(s_1, \ldots, s_5)$ and two slack variables (t_1, t_2):

$$\begin{array}{ccccc} \min u + v + w & \text{subject to} & x_1, x_2, y_1, y_2, s_1, s_2, s_3, s_4, s_5, t_1, t_2 \geq 0 & \text{and} & \\ u - x_1 + x_2 - s_1 = 0 & v - y_1 + y_2 - s_3 = 0 & & w + 2x_1 - 2x_2 - s_5 = 3 & \\ u + x_1 - x_2 - s_2 = 0 & v + y_1 - y_2 - s_4 = 0 & & 2x_1 - 2x_2 - w + t_1 = 3 & \\ & & x_1 - x_2 + y_1 - y_2 + t_2 = 1. & & \end{array} \tag{3.3}$$

The nonnegativity constraints on u, v and w are still missing. Observe, however, that the two constraints involving u sum up to $2u - s_1 - s_2 = 0$ and together with $s_i \geq 0$ they imply that $u \geq 0$. Therefore, we can harmlessly add the constraint that $u \geq 0$; it is implied by the other constraints. A similar argument applies to v and w (for w one needs to subtract the second constraint from the first). Once we add these constraints, (3.3) takes its standard form, with fourteen variables and seven constraints.

The reason we postponed the observation about positivity of u, v and w is that considering them as free variables allows simplifying the program (3.3). Indeed, we can use the three

equalities

$$u = x_1 - x_2 + s_1 \qquad v = y_1 - y_2 + s_3 \qquad w = 3 - 2x_1 + 2x_2 + s_5$$

to write (3.3) as

$$\begin{array}{l} \min 3 + x_2 - x_1 + y_1 - y_2 + s_1 + s_3 + s_5 \qquad \text{subject to} \quad x_1, x_2, y_1, y_2, s, t \geq 0 \\ \qquad x_1 - x_2 + y_1 - y_2 + t_2 = 1 \quad 2x_1 - 2x_2 + s_1 - s_2 = 0 \\ \qquad 2y_1 - 2y_2 + s_3 - s_4 = 0 \qquad 4x_1 - 4x_2 - s_5 + t_1 = 6. \end{array} \tag{3.4}$$

Ignoring the additive constant 3, (3.4) is a linear program in standard form, with eleven variables and four constraints. To express it in matrix form, write

$$\mathbf{x} = (x_1, x_2, y_1, y_2, s_1, s_3, s_5, s_2, s_4, t_1, t_2)^t.$$

(The variables s_1, s_3 and s_5 are no longer surplus variables, because they now appear in the objective function; hence this peculiar ordering.) Then (3.4) writes

$$\min(-1,1,1,-1,1,1,1,0,0,0,0)\mathbf{x} \qquad \text{subject to} \quad \mathbf{x} \geq 0 \quad \text{and}$$

$$\left(\begin{array}{ccccccc|cccc} 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 2 & -2 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 4 & -4 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \end{array}\right) \mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 6 \end{pmatrix}.$$

The vertical line separates the slack and surplus variables (s_2, s_4, t_1, t_2) from the other variables, and serves for convenience only. Once this problem is solved, the original variables are given by $x = x_1 - x_2$, $y = y_1 - y_2$ and $z = 3 - 2x$.

Remark 3 *The problem (3.4) can in fact be simplified even further by replacing $x = x_1 - x_2$, $y = y_1 - y_2$, making x and y free and using two of the constraints to eliminate them. The simplified linear program will have seven variables and two constraints only.*

Another example of a nonlinear objective function is

$$\min \max(c^t x + \alpha, d^t x + \beta) \qquad \text{subject to} \qquad Ax = b \quad \text{and} \quad x \geq 0, \tag{3.5}$$

where c, d and x are in $\mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$. (3.5) can be reduced to a linear program by an argument similar to the one in the preceding example: introduce the new variable u, which should replace the maximum. Add the constraints $u \geq c^t x + \alpha$ and $u \geq d^t x + \beta$, and consider the program

$$\begin{array}{ccccc} & & \min u & \text{subject to} & \\ Ax = b & & c^t x - u \leq -\alpha & d^t x - u \leq -\beta \quad \text{and} \quad x \geq 0. & \end{array} \tag{3.6}$$

(3.6) is a linear program, to which any feasible solution satisfies $u \geq \max(c^t x + \alpha, d^t x + \beta)$. As these are the only constraints on u and it is to be minimized, it is clear that in any optimal solution $u = \max(c^t x + \alpha, d^t x + \beta)$. Hence the equivalence between (3.6) and the original problem (3.5). The transformation to standard form is then carried out as in the previous example. This example is more general than (3.1), because $|x| = \max(x, -x)$.

3.3 Basic solutions

In this subsection we present the fundamental notion of linear programming. Since no loss of generality results by dealing with linear programs in standard form, we only treat problems of this form.
A linear dependence between the rows of A means that either the constraints entail a contradiction, or that some of them can are redundant. We therefore assume from now on that the m rows of A are linearly independent. This condition is sometimes referred to as *constraint qualification*. It follows immediately that there are at least m columns, so that $n \geq m$. In words, there are at least as many variables as constraints. This ensures that the set $\{x : Ax = b\}$ is nonempty: let $i_1 < i_2 < \cdots < i_m$ be the indices of m independent columns and form a matrix $B \in M_{m,m}$ from these columns, then the matrix B is invertible. Let $x^B = B^{-1}b \in \mathbb{R}^m$, and augment x^B with zeros so that the augmented vector will lie in $\mathbb{R}^n$; define

$$x = (x_1, \ldots, x_n) \qquad \text{where} \quad x_k = \begin{cases} x_j^B & \text{if } k = i_j \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

Then $Ax = Bx^B = b$.
We often abuse notation and say that $\{i_1, \ldots, i_m\}$ is a basis, meaning that the columns $\{A_{i_1}, \ldots, A_{i_m}\}$ are a basis of $\mathbb{R}^m$.

Definition 3.2 *A solution $x \in \mathbb{R}^n$ of the system $Ax = b$ is **basic** if it is of the form 3.7. Equivalently, the collection of columns of A*

$$\{A_i : x_i \neq 0\}$$

*is a linearly independent set (in $\mathbb{R}^m$). x is **degenerate** if less than m of its coordinates are nonzero, otherwise it is **non degenerate**.*

If x is a non degenerate basic solution, then the basis B is determined as the set of indices of the nonzero coordinates of x. If x is degenerate, however, then this is impossible; the basis is not uniquely defined since any collection of m independent columns that includes those corresponding to the nonzero values of x forms a basis corresponding to x.
Given a basis B, the corresponding variables are said to be **basic**, while the other variables are **nonbasic**.
While the set $\{x : Ax = b\}$ is nonempty, the set $\{x : Ax = b, x \geq 0\}$ can of course be empty: for example, no combination of nonnegative numbers can satisfy $x_1 - x_2 = 5$ and $x_1 + x_3 = 2$. In words, a basic solution may fail to be feasible.

Remark 4 *If $b = 0$ then the only basic solution is the zero solution (which is feasible), because $B^{-1}0 = 0$ for any nonsingular B. Other solutions may exist, however, for example if $A = \begin{pmatrix} 1 & -1 \end{pmatrix} \in M_{1,2}$.*

There are only finitely many basic solutions: indeed, each one corresponds to the selection of m independent columns from A. Since A has n columns, there are at most

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

distinct selections. This observation, together with the following theorem, provide the basis for the simplex algorithm.

Theorem 3.3 (The fundamental theorem of linear programming) *Consider the linear program*

$$\min c^t x \qquad \textit{subject to} \quad Ax = b \quad \textit{and} \quad x \geq 0,$$

where $A \in M_{m,n}$ is of rank m. Then
(1) if a feasible solution exists, then a feasible basic solution exists.
(2) if an optimal solution exists, then an optimal basic solution exists.

Proof. (1) Let x be a feasible solution, and assume that it has k nonzero coordinates, $i_1 < \cdots < i_k$. Let B be the matrix formed by the columns A_{i_j}, $j = 1, \ldots, k$. If these columns are linearly independent, then x is basic and the claim holds trivially. Otherwise there exist some constants $a_1, \ldots, a_k \in \mathbb{R}$ such that $a_1 A_{i_1} + \cdots + a_k A_{i_k} = \mathbf{0} \in \mathbb{R}^m$ and $\max a_i > 0$. Augment the vector $(a_1, \ldots, a_k)$ as in (3.7) by defining

$$y = (y_1, \ldots, y_n) \qquad \text{where} \quad y_s = \begin{cases} a_j & s = i_j \\ 0 & \text{otherwise.} \end{cases}$$

Then $A(\varepsilon y) = \varepsilon A y = 0$ for any $\varepsilon \in \mathbb{R}$, and $y \neq 0$. Whenever $x_i = 0$, $y_i = 0$ too, so that $x - \varepsilon y$ is a solution having at most k nonzero coordinates. x is feasible, that is, nonnegative, and we look for a value of ε so that $x - \varepsilon y$ be feasible as well. Choose $\varepsilon = \min\{x_j/y_j : y_j > 0\}$, then $\varepsilon > 0$ because if $x_j = 0$ then $y_j = 0$. Then $x - \varepsilon y$ is a feasible solution with at most $k - 1$ nonzero coordinates (because for at least one j, $0 \neq x_j = \varepsilon y_j$). We then replace x by $x - \varepsilon y$ and continue this procedure until the a basic feasible solution is obtained.
(2) Let x be again as in the proof of (1) and suppose furthermore that it is optimal. Then the same proof of (1) works, except that we need to show that $x - \varepsilon y$ is optimal too. As $c^t(x - \varepsilon y) = c^t x - \varepsilon c^t y$ it is sufficient to show that $c^t y = 0$. Let $\rho = \min\{|x_j/y_j| : y_j \neq 0\}$ then $\rho > 0$ by the same argument for ε and $x + \rho y$ is a feasible solution. If $c^t y < 0$ then $c^t(x + \rho y) < c^t x$; if $c^t y > 0$ then $c^t(x - \varepsilon y) < c^t x$. Since both cases contradict the optimality of x we conclude that $c^t y = 0$.

Remark 5 *Replacing c by $-c$, we obtain that the worst value, if exists, is also attained by a basic feasible solution. That is, there exists a basic feasible solution y such that*

$$c^t y = \sup\{c^t x : Ax = b, x \geq 0\}$$

provided the right hand side exists and is finite.

The fundamental theorem of linear programming gives rise to a simple algorithm: choose all combinations of m independent columns, find the basic solution corresponding to each of them (e.g. by inverting a matrix) and compare all the results. This procedure, though highly inefficient, is guaranteed to be finite, as there are finitely many basic solutions. The best result that is attained from a feasible solution corresponds to an optimal solution, provided that one exists.
The latter condition is essential. As an example of a case where no optimal solutions exist, consider the linear program

$$\begin{array}{lll} \min x - y + z & & \text{subject to} \\ 2x - z = 2 & 3x - y + 3z = 6 & x, y, z \geq 0. \end{array}$$

There are three variables and two (independent) constraints. To find the basic solutions, we set each of the variables in turn to zero and then determine the other two from the constraints. One verifies that the basic solutions are $(0, -12, -2)$, $(4/3, 0, 2/3)$ and $(1, -3, 0)$, of which only the second is feasible. Therefore, if there is an optimal solution, the objective function must attain the value of $4/3 - 0 + 2/3 = 2$. However, $(2, 6, 2)$ is also feasible and attains the smaller value -2. From this and Theorem 3.3 it follows that no optimal solution exists. This can also be seen directly by substituting $z = 2x - 2$ and $y = 3x - 6 + 3z = 9x - 12$, then the problem is formulated in terms of the variable x only:

$$\min 10 - 6x \qquad \text{subject to} \quad 2x - 2 \geq 0 \quad 9x - 12 \geq 0 \quad x \geq 0.$$

Or equivalently,

$$\min 10 - 6x \qquad \text{subject to} \quad x \geq \frac{4}{3}.$$

The objective function decreases with x, which is unbounded from above. Therefore no optimal solution exists; the set of attainable values is unbounded from below. The feasible solution $x = 4/3$ corresponds to the only basic feasible solution of the program, yielding a value of 2 for the objective function. This is indeed the worst feasible solution of the program as in Remark 5.
When the objective value can attain arbitrary large negative values, we say that the program is **unbounded**. This implies that the set of feasible solutions is unbounded, but the converse does not necessarily hold.

3.4 Relation to convexity

For any $a \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$, the set $\{x : a^t x = \beta\}$ is convex and closed. We therefore have immediately that

Lemma 3.4 *Let* $\min c^t x$ *subject to* $Ax = b$ *and* $x \geq 0$ *be a linear program. Then the set of feasible solutions of the program is convex and closed. The same holds for the set of optimal feasible solution.*

Proof. The intersection of arbitrarily many convex sets is convex, and a same relation holds for closed sets. Since $\{x \in \mathbb{R}^n : x \geq 0\}$ is closed and convex, the first assertion follows from the observation before the lemma. If an optimal solution y exists, then the set of optimal solutions is $\{x \in \mathbb{R}^n : Ax = b, x \geq 0, c^t x = c^t y\}$, a convex closed set. Noting that the empty set is both closed and convex, the proof is complete.

Definition 3.5 *A **polytope** is a set of the form*

$$\{x \in \mathbb{R}^n : Ax = b, \tilde{A}x \leq \tilde{b}\} \qquad A \in M_{m,n} \quad b \in \mathbb{R}^m_+ \quad \tilde{A} \in M_{k,n} \quad \tilde{b} \in \mathbb{R}^k.$$

*A bounded polytope is said to be a **polyhedron**.*

For a linear program, the set of feasible solutions is a polytope. When it is given in standard form, the only inequality constraints are $x \geq 0$, so that $k = n$, $\tilde{A} = -I_n$ and $\tilde{b} = 0 \in \mathbb{R}^n$.

Definition 3.6 *Let K be a convex set. $x \in K$ is an **extreme point** of K if whenever $x = tx_1 + (1-t)x_2$ for $t \in (0,1)$ and $x_1, x_2 \in K$, one has $x_1 = x_2$.*

A linear function on a convex set attains its extrema at extreme points of the set, and therefore (at least for bounded polytopes) we can consider the extreme points only for the minimization problem. The following proposition identifies the extreme points of K with the basic solutions, providing an alternative proof for the fundamental theorem.

Proposition 3.7 *Let K be a polytope defined by the constraints $Ax = b$ and $x \geq 0$. The set of extreme points of K is the set of the basic feasible solutions.*

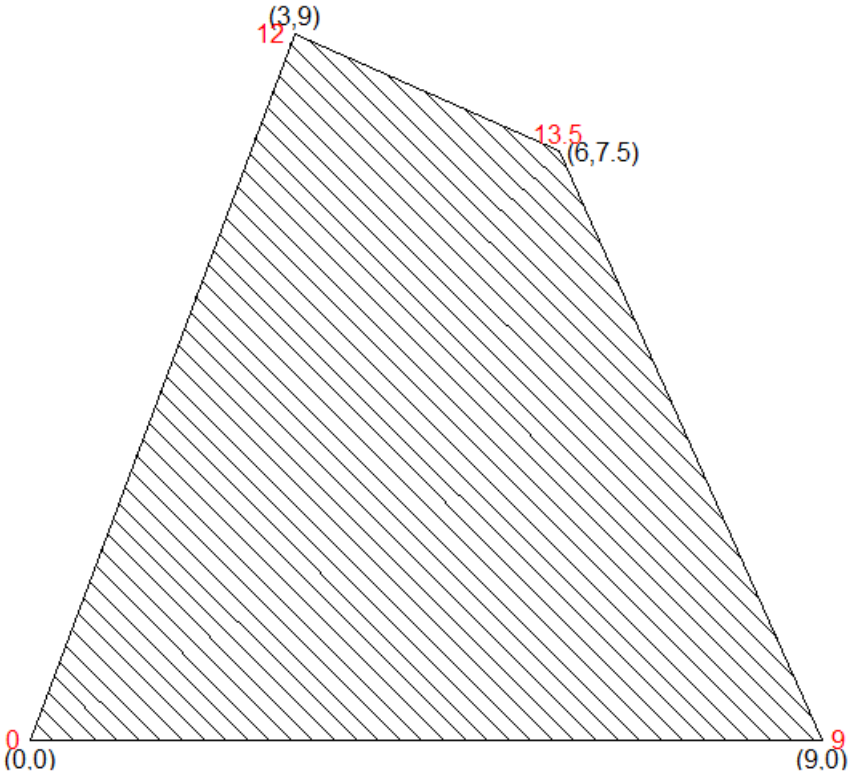
Proof. Let x be a feasible solution, and assume for simplicity that its nonzero entries are $x_1, \ldots, x_p$.
Suppose that x is nonbasic. Let $B \in M_{m,p}$ be the matrix formed by the first p columns of A, then the equation $Bz = 0$ for $z \in \mathbb{R}^p$ has infinitely many solutions. Choose a nonzero solution z and scale it so that $|z_i| \leq x_i$ for $i = 1, \ldots, p$. Consider $y = (z, 0) \in \mathbb{R}^n$ where $0 \in \mathbb{R}^{n-p}$, then $Ay = Bz = 0$. Both $x + y$ and $x - y$ are nonnegative, and distinct since $y \neq 0$. They are feasible solutions because $Ay = 0$. The equality $x = (x+y)/2 + (x-y)/2$ means that x is not an extreme point of K.
Suppose now that x is basic, and that $x = ty + (1-t)z$ for $t \in (0,1)$ and $y, z \in K$. If $z_k > 0$ for $k > p$, then $y_k = (t-1)z_k/t < 0$, which is impossible since y is feasible. Therefore only $z_1, \ldots, z_p$ can be nonzero. This means that $x - z$ is a basic solution to the system of equations $Au = 0$, which is only possible if $x = z$ by Remark 4. It follows that $x = z = y$ and x is an extreme point.

In many cases in practice, the polytope K is bounded and nonempty, hence it is a compact set. The continuous function $x \mapsto c^t x$ obtains a minimum on K, and this minimum is

obtained at a basic solution which is an extreme point of K. This fails to hold if K is unbounded. K is guaranteed to be bounded if some row of A contains only strictly positive entries. An example of a bounded polytope is given in Figure 1. It corresponds to the linear program

$$\begin{array}{lll} \max x+y & \text{subject to} & x,y \geq 0 \quad \text{and} \\ y-3x \leq 0 & 2y+x \leq 21 & 2y+5x \leq 45. \end{array} \tag{3.8}$$

Figure 1: The polytope determined by the feasible points of (3.8), its extreme points and the objective values at these points



The lines inside the polytope are the contours of the objective function $f(x,y) = x+y$. It is clear from the figure that both extrema of f are attained at one of the four extreme points of the polytope.

3.5 Moving to a better basic solution

In the preceding subsections it is shown how solving a linear program reduces to the comparison of the objective function at the (finitely many) basic solutions. There can be at most $n!/(m!(n-m)!)$ basic solutions, since any of these is determined by choosing m independent columns from A. The simplex algorithm is based on this idea: it jumps from one basic solution to another while decreasing the objective value at each step.

Furthermore, it does so in a way that requires row manipulations only at each step, which require substantially less computational effort than inverting a matrix.
Consider the program

$$\min c^t x \qquad \text{subject to} \quad Ax = b \quad \text{and} \quad x \geq 0.$$

Suppose that first m columns of A form a basis, and that $x = (x_1, \ldots, x_m, 0, \ldots, 0)^t$ is a basic feasible solution, with objective value $\sum_{i=1}^m c_i x_i$. Now consider introducing the variable x_q into the basis, for $q > m$. How can this be done? The columns $A_1, \ldots, A_m$ span $\mathbb{R}^m$, so that $A_q = \sum_{i=1}^m \alpha_{iq} A_i$. Choosing $y = (-\alpha_{1q}, \ldots, -\alpha_{mq}, 0, \ldots, 0, 1, 0, \ldots, 0)^t$ (where the value one is at the q-th coordinate), we have $Ay = 0$ and therefore $x + \varepsilon y$ is a solution for any $\varepsilon \in \mathbb{R}$. The questions to ask now are:
(1) Is $x + \varepsilon y$ feasible?
(2) Is $x + \varepsilon y$ better than x? In other words, is the value of the objective function at $x + \varepsilon y$ lower than the one at x?
For $\varepsilon < 0$, $x + \varepsilon y$ is infeasible, because its q-th coordinate is $x_q + \varepsilon y_q = \varepsilon$. For $\varepsilon > 0$, we need to verify that $x_i - \varepsilon \alpha_{iq} \geq 0$ for any $i \leq m$, and such $\varepsilon > 0$ exists if and only if the following relation holds: whenever $\alpha_{iq} > 0$, $x_i > 0$ as well. In particular, if x is non degenerate, then ε must exist. The maximal ε maintaining feasibility is $\varepsilon^* \stackrel{\text{def}}{=} \min\{x_i/\alpha_{iq} : \alpha_{iq} > 0\}$, which is by convention infinite if $\alpha_{iq} \leq 0$ for any i. In this latter case the set of feasible solutions is unbounded.
The question whether $x + \varepsilon y$ is better than x is equivalent to whether $\varepsilon c^t y$ is negative or not. Since only positive values of ε can lead to a feasible solution, we are simply interested in the sign of $c^t y$. But

$$c^t y = c_q - \sum_{i=1}^m \alpha_{iq} c_i = c_q - z_q,$$

where z_q is defined by the above equality. It is fruitful to introduce the variable x_q into the basis if and only if $c_q < z_q$, in which case we wish to use the largest value of ε that maintains feasibility, i.e. ε^*. If ε^* is finite, then it equals x_p/α_{pq} for some $p \leq m$. Then $x_p + \varepsilon^* y_p = 0$, and the (better) solution $x + \varepsilon^* y$ has at most m nonzero coordinates — $([m] \cup \{q\}) \setminus \{p\}$, where $[m] = \{1, \ldots, m\}$. The corresponding columns span $\mathbb{R}^m$, because

$$A_p = \frac{1}{\alpha_{pq}} A_q - \sum_{i \neq p} \frac{\alpha_{iq}}{\alpha_{pq}} A_i.$$

This means that $x + \varepsilon^* y$ too is a basic feasible solution.
If $\varepsilon^* = \infty$ and $c_q < z_q$ then $x + \varepsilon y$ is a feasible solution for any $\varepsilon \geq 0$, and the objective function can be decreased without bound, so that there is no optimal solution for the program. If $\varepsilon^* = \infty$ and $c_q \geq z_q$ then the set of feasible solutions is unbounded but an optimal solution might or might not exist.
If $\varepsilon^* = 0$ then $x + \varepsilon^* y = x$; the basic solution does not change, but the basis changes from $[m]$ to $([m] \cup \{q\}) \setminus \{p\}$.

It is convenient to define $r_j = c_j - z_j$. Introducing the variable x_j into the basis is advantageous if and only if $r_j < 0$. r_j is called the **relative cost coefficient** of the variable x_j (with respect to the current basis). For $j \leq m$ we have $\alpha_{ij} = \delta_{ij}$ and $r_j = c_j - c_j = 0$, so that the relative cost coefficient of any basic variable is zero. For an economical interpretation of the relative cost coefficients, see [9, p. 45–46].

For a geometrical visualization of these abstract considerations, we reconsider the linear program (3.8) from Figure 1, which we write in standard form by adding slack variables

$$\min -x - y \qquad \text{subject to} \quad x, y, s \geq 0 \quad \text{and}$$
$$\begin{pmatrix} -3 & 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 \\ 5 & 2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 21 \\ 45 \end{pmatrix}. \tag{3.9}$$

(We multiplied the objective function by -1 in order to have a minimization problem.) Geometrically, the slack variables measure the vertical distance of each point from the line that represents their corresponding constraint. For example, for the point $x = 3$ and $y = 9$, $s_1 = s_2 = 0$ because this point is on the intersection of the lines of the constraints $y \leq 3x$ and $x + 2y \leq 21$; these two constraints are satisfied with equality. $s_3 = 12$ because strict inequality holds for the constraint $5x + 2y \leq 45$. In fact $5x + 2y = 45 - 12 = 45 - s_3$. A natural way to form a basis, that is, three independent columns, is to choose the last three columns, which form the identity matrix. This choice corresponds to the degenerate basic solution

$$d = (0, 0, 0, 21, 45),$$

i.e. $x = y = 0$ and $s = (0, 21, 45)$. Denoting by A the 3×5 constraint matrix, it is easily seen that $\alpha_{iq} = A_{iq}$, rendering the subsequent calculations easier. Since the costs of the slacks are zero, we see that $r_x = c_x = -1$ and $r_y = c_y = -1$, so introducing any of them to the basis improves the objective. To introduce x we see that the values of α_{iq} are -3, 1 and 5 hence the vector

$$z = (1, 0, 3, -1, -5)$$

satisfies $Az = 0$. Thus, any vector of the form

$$d + \varepsilon z = (\varepsilon, 0, 3\varepsilon, 21 - \varepsilon, 45 - 5\varepsilon)$$

is a solution, and it is feasible if and only if $\varepsilon \in [0, 9]$. 9 is indeed the minimal ratio among 45/5 and 21/1, where the ratio 0/-3 is not taken into account because -3 is negative. Geometrically, the choice of $\varepsilon \in [0, 9]$ is equivalent to the choice of a point on the corresponding edge of the polytope in Figure 1. As we move along this edge in the positive direction, the objective value decreases, and the maximal decrease under feasibility is obtained when $\varepsilon = 9$, leading to the basic non degenerate solution

$$(9, 0, 27, 12, 0). \tag{3.10}$$

The new basis is formed by the first, third and fourth columns of the matrix A: x became basic, and s_3 became nonbasic.

We now consider what would have happened if we entered the variable y instead of x. In this case, we have

$$z = (0, 1, -1, -2, -2),$$

and

$$d + \varepsilon z = (0, \varepsilon, -\varepsilon, 21 - 2\varepsilon, 45 - 2\varepsilon).$$

We see that the only feasible value of ε is 0, due to the minus sign in the third coordinate. Thus, y can become basic instead of s_1, but the basic solution remains the same. This algebraic observation can also be seen in Figure 1, where $y > 0$ is only possible when $x > 0$.
Once y is in the basis, we can recalculate the relative cost coefficients of the nonbasic variables x and s_1. Our current basis is given by the second, fourth and fifth columns and we need to find the representations of the other columns as a linear combination of the basis

$$(-3, 1, 5) = -3(1, 2, 2) + 7(0, 1, 0) + 11(0, 0, 1) \Rightarrow r_x = c_x + 3c_y - 7c_{s_2} - 11c_{s_3} = -4$$

and

$$(1, 0, 0) = (1, 2, 2) - 2(0, 1, 0) - 2(0, 0, 1) \Rightarrow r_{s_1} = c_{s_1} - c_y + 2c_{s_2} + 2c_{s_3} = 1.$$

We see that it is profitable to introduce x into the basis; we have

$$z = (1, 3, 0, -7, -11) \Rightarrow d + \varepsilon z = (\varepsilon, 3\varepsilon, 0, 21 - 7\varepsilon, 45 - 11\varepsilon),$$

which is feasible for $\varepsilon \in [0, 3]$. We move along the leftmost edge of the polygon, where $y = 3x$ and $s_1 = 0$; we do this until arrival to $(3, 9, 0, 0, 12)$, and so on.

This example illustrates the basic procedure of the simplex algorithm. Given a basis B, one calculates the relative cost coefficients of the nonbasic variables. If one of them is negative, introduce this variable to the basis as in the example. If the move is non degenerate, this amounts to moving along an edge of the polytope of feasible solutions until another vertex is reached; the objective function decreases as we advance. If no vertex is reached, then the problem is unbounded. Otherwise, repeat this procedure with the new basis.
Assuming that any basic solution is non degenerate, it follows that in each step the objective function decreases by a positive amount. Hence the algorithm can never return to a point it visited before. As there are finitely many basic solutions, the procedure must terminate at a point for which none of the relative cost coefficients is negative (or in

determining that the problem is unbounded). The lemma below proves that this condition ensures optimality and also a unveils a way to calculate the relative cost coefficients. Later on we deal with the existence of degenerate solutions.

Lemma 3.8 *Let $d = (d_1, \dots, d_m, 0, \dots, 0)$ be a basic feasible solution and write*

$$A_q = \sum_{i=1}^{m} \alpha_{iq} A_i \qquad m+1 \leq q \leq n.$$

Define $r_q = c_q - \sum_{i=1}^{m} \alpha_{iq} c_i$ for $q = m+1, \dots, n$ and suppose that $r_q \geq 0$ for any q. Then d is optimal. If the inequality is strict for any q then d is uniquely optimal.

Proof. The first m columns of A are independent by assumption, so after elementary manipulations on the rows of the matrix A, it can be brought to the form $B = [I|\beta]$. This means that the constraints $Ax = b$ are equivalent to

$$Bx = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{pmatrix}, \qquad B \stackrel{\text{def}}{=} \begin{pmatrix} 1 & & & \beta_{1,m+1} & \dots & \beta_{1q} & \dots & \beta_{1n} \\ & 1 & & \beta_{2,m+1} & \dots & \beta_{2q} & \dots & \beta_{2n} \\ & & \ddots & \vdots & & \vdots & & \\ & & 1 & \beta_{m,m+1} & \dots & \beta_{mq} & \dots & \beta_{mn} \end{pmatrix}. \tag{3.11}$$

Since we only performed linear maps on the rows, the column relation

$$A_q = \sum_{i=1}^{m} \alpha_{iq} A_i \qquad m+1 \leq q \leq n$$

still holds for B, which means that $\beta_{ij} = \alpha_{ij}$ for any i, j.
The discussion at the beginning of the subsection shows that no gain can be made by introducing a single variable x_q ($q > m$) into the basis, but it is not a-priori clear that no such gain can be made by introducing several variables simultaneously. To show that this is the case, we observe that given the values of $x_{m+1}, \dots, x_n$, the values of $x_1, \dots, x_n$ are determined by the constraints (3.11):

$$x_i = d_i - \sum_{j=m+1}^{n} \beta_{ij} x_j = d_i - \sum_{j=m+1}^{n} \alpha_{ij} x_j \qquad i = 1, \dots, m.$$

The objective function can then be written as a function of the nonbasic variables only

$$c^t x = \sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{m} c_i \left(d_i - \sum_{j=m+1}^{n} \alpha_{ij} x_j \right) + \sum_{j=m+1}^{n} c_j x_j =$$

$$= c^t d + \sum_{j=m+1}^{n} \left(c_j - \sum_{i=1}^{m} \alpha_{ij} c_i \right) x_j = c^t d + \sum_{j=m+1}^{n} r_j x_j \geq c^t d,$$

since $r_j \geq 0$ by the hypothesis and $x_j \geq 0$ for any feasible solution. This means that the objective function at any feasible point attains a value at least as large as the one attained

at d; d is therefore optimal. If $r_j > 0$ for any j then the objective function strictly exceeds $c^t d$ unless all the nonbasic variables vanish, so d is the unique optimal solution.

From the proof of Lemma 3.8 we see that the coefficients α_{iq} are precisely the elements of column q, provided that the columns corresponding to the basic variables in the constraints matrix are the standard unit vectors in $\mathbb{R}^m$. In order to enter q into the basis, we apply row manipulations on the constraint matrix so that the q-th column becomes one of the standard unit vectors. During this process, one column that was a standard unit vector becomes an arbitrary vector; its corresponding variable becomes nonbasic. If the current solution is d, the variable to leave the basis bears the index i for which $\alpha_{iq} > 0$ and the ratio d_i/α_{iq} is minimal.

To see this in an alternative way, observe that the same row manipulations applied to the constraints matrix have to be carried out on the vector b as well. If the basis is the identity matrix, then $b = d$. When changing the basis d changes as well, and we must make sure that it remains nonnegative. The only way to do this is to remove the variable minimizing d_i/α_{iq} from the basis.

To illustrate this discussion, consider again the program (3.9). Attaching the vector b as an additional column, we obtain

$$\begin{pmatrix} -3 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 21 \\ 5 & 2 & 0 & 0 & 1 & 45 \end{pmatrix}, \tag{3.12}$$

where the columns correspond to the variables x, y, s_1, s_2 and s_3. The last three columns form the identity matrix and the basic solution is $(0, 0, 0, 21, 45)$, which is degenerate. If we wish to enter x into the basis, the variable that has to leave is the one with the minimal ratio between $21/1$ and $45/5$. The ratio is minimal in the third row, and so s_3 has to leave the basis. To make the first column equal $(0, 0, 1)^t$ we use the following row manipulations: A^1 is replaced by $A^1 + (3/5)A^3$, A^2 is replaced by $A^2 - (1/5)A^3$ and A^3 is replaced by $A^3/5$. The reader can notice without any calculation that the columns of the other basic variables, s_1 and s_2, remain unchanged. The matrix then becomes

$$\begin{pmatrix} 0 & 11/5 & 1 & 0 & 3/5 & 27 \\ 0 & 8/5 & 0 & 1 & -1/5 & 12 \\ 1 & 2/5 & 0 & 0 & 1/5 & 9 \end{pmatrix}, \tag{3.13}$$

yielding the basic solution $(9, 0, 27, 12, 0)$ (as in (3.10)). Had we (incorrectly) chosen to remove s_2 from the basis in (3.12), the manipulations to carry out are to replace A^1 by $A^1 + 3A^2$, A^3 by $A^3 - 5A^2$ and leave A^2 unchanged. The matrix would become

$$\begin{pmatrix} 0 & 7 & 1 & 3 & 0 & 63 \\ 1 & 2 & 0 & 1 & 0 & 21 \\ 0 & -8 & 0 & -5 & 1 & -60 \end{pmatrix},$$

leading to the infeasible basic solution $(21, 0, 63, 0, -60)$. It corresponds to the point $(21, 0)$, where the line $x + 2y = 21$ intersects the x axis.

As discussed above, the relative cost coefficients in (3.13) are easily calculated. Recalling that the objective vector is

$$c = (-1, -1, 0, 0, 0) = (c_x, c_y, c_{s_1}, c_{s_2}, c_{s_3}),$$

we have

$$r_y = c_y - (11/5)c_{s_1} - (8/5)c_{s_2} - (2/5)c_x = -1 + 2/5 = -3/5 < 0$$

and

$$r_{s_3} = c_{s_3} - (3/5)c_{s_1} + (1/5)c_{s_2} - (1/5)c_x = 1/5.$$

It follows that y should be entered into the basis. All the coefficients in its column are positive, and the minimal ratio is $12/(8/5)$. Thus the variable to leave the basis is s_2. The manipulations are adding multiples of the second row to the other rows and normalizing the second row:

$$\begin{pmatrix} 0 & 0 & 1 & -11/8 & 7/8 & 10.5 \\ 0 & 1 & 0 & 5/8 & -1/8 & 7.5 \\ 1 & 0 & 0 & -1/4 & 1/4 & 6 \end{pmatrix}. \tag{3.14}$$

This corresponds to the basic solution $(6, 7.5, 10.5, 0, 0)$. We have

$$r_{s_2} = (11/8)c_{s_1} - (5/8)c_y + (1/4)c_x = 3/8 \qquad r_{s_3} = (-7/8)c_{s_1} + (1/8)c_y - (1/4)c_x = 1/8. \tag{3.15}$$

As all the relative cost coefficients are positive, this solution is uniquely optimal by Lemma 3.8.

3.6 The simplex tableau and the algorithm

There is an even simpler method for calculating the relative cost coefficients. We have seen before that the vector b can be attached to the constraint matrix A as another column; as we shall see, the relative cost coefficients can be added as an extra row, and this row can be treated like any other row when applying the row manipulations.
The linear program in standard form

$$\min c^t x \qquad \text{subject to} \quad Ax = b, \quad x \geq 0 \tag{3.16}$$

is equivalent to

$$\min z \qquad \text{subject to} \quad Ax = b, \quad c^t x - z = 0, \quad x \geq 0, \quad z \text{ free}. \tag{3.17}$$

The constraints matrix of (3.17) with its right hand side vector $(b, 0)$ takes the form

$$A_c^b = \begin{pmatrix} A_{11} & \dots & A_{1q} & \dots & A_{1n} & 0 & b_1 \\ A_{21} & \dots & A_{2q} & \dots & A_{2n} & 0 & b_2 \\ \vdots & & \vdots & & \vdots & \vdots & \vdots \\ A_{m1} & \dots & A_{mq} & \dots & A_{mn} & 0 & b_m \\ c_1 & \dots & c_q & \dots & c_n & -1 & 0 \end{pmatrix}. \tag{3.18}$$

The $(m+1)\times(n+1)$ matrix A_c^b contains all the information about the linear program. It is called **the simplex tableau**. Now, given any basis B of the original problem (3.16) and its corresponding basic solution d, the last constraint $c^t x = z$ is equivalent to $z = c^t d + \sum_{j\notin B} r_j x_j$, or $\sum_{j\notin B} x_j r_j - z = -z_0$, where $z_0 = c^t d$ is the objective value at d and the r_j's are the relative cost coefficients with respect to the basis B. If the basis is given by the first m columns then row manipulations on (3.18) lead to

$$\begin{pmatrix} 1 & & & \alpha_{1,m+1} & \dots & \alpha_{1q} & \dots & \alpha_{1n} & 0 & d_1 \\ & 1 & & \alpha_{2,m+1} & \dots & \alpha_{2q} & \dots & \alpha_{2n} & 0 & d_2 \\ & & \ddots & \vdots & & \vdots & & \vdots & \vdots & \vdots \\ & & & 1 & \alpha_{m,m+1} & \dots & \alpha_{mq} & \dots & \alpha_{mn} & 0 & d_m \\ c_1 & & \dots & c_m & c_{m+1} & \dots & c_q & \dots & c_n & -1 & 0 \end{pmatrix}.$$

Finally, subtracting from the last row c_1 times the first row, c_2 times the second and so on the simplex tableau becomes

$$R = \begin{pmatrix} 1 & & & \alpha_{1,m+1} & \dots & \alpha_{1q} & \dots & \alpha_{1n} & d_1 \\ & 1 & & \alpha_{2,m+1} & \dots & \alpha_{2q} & \dots & \alpha_{2n} & d_2 \\ & & \ddots & \vdots & & \vdots & & \vdots & \vdots \\ & & & 1 & \alpha_{m,m+1} & \dots & \alpha_{mq} & \dots & \alpha_{mn} & d_m \\ 0 & & \dots & 0 & r_{m+1} & \dots & r_q & \dots & r_n & -z_0 \end{pmatrix}, \quad z_0 = c_1 d_1 + c_2 d_2 + \cdots + c_m d_m, \tag{3.19}$$

where we have omitted the column $(0,\dots,0,-1)$ because it never changes during the process. Thus, we see that the relative cost coefficients appear on the last row, and can be derived by elementary row manipulations. When the basis is formed by different columns, the form of the simplex tableau is as in (3.19) up to a permutation of the columns (except the last one). It is then said to be in **canonical form**.

Suppose that in (3.19) $r_{m+1} < 0$ and we wish the variable x_{m+1} to enter the basis. Suppose that the minimal ratio $d_i/\alpha_{i,m+1}$ over the positive $\alpha_{\cdot,m+1}$ is $d_2/\alpha_{2,m+1}$. Then x_2 should leave the basis and the $(m+1)$-th column should become $(0,1,0,\dots,0)^t$. To do this we replace the i-th row R^i by $R^i - (\alpha_{i,m+1}/\alpha_{2,m+1})R^2$ for $i \neq 2$ and R^2 by $R^2/\alpha_{2,m+1}$. Let $\alpha = \alpha_{2,m+1} > 0$ then the resulting simplex tableau

$$\begin{pmatrix} 1 & -\frac{\alpha_{1,m+1}}{\alpha} & & & 0 & \dots & \alpha_{1q} - \frac{\alpha_{1,m+1}}{\alpha}\alpha_{2q} & \dots & \alpha_{1n} - \frac{\alpha_{1,m+1}}{\alpha}\alpha_{2n} & d_1 - \frac{\alpha_{1,m+1}}{\alpha}d_2 \\ & 1/\alpha & & & 1 & \dots & \alpha_{2q}/\alpha & \dots & \alpha_{2n}/\alpha & d_2/\alpha \\ & -\frac{\alpha_{3,m+1}}{\alpha} & 1 & & 0 & & \alpha_{3q} - \frac{\alpha_{3,m+1}}{\alpha}\alpha_{2q} & & \alpha_{3n} - \frac{\alpha_{3,m+1}}{\alpha}\alpha_{2n} & d_3 - \frac{\alpha_{3,m+1}}{\alpha}d_2 \\ & \vdots & & \ddots & \vdots & & \vdots & & \vdots & \vdots \\ & -\frac{\alpha_{m,m+1}}{\alpha} & & & 1 & 0 & \dots & \alpha_{mq} - \frac{\alpha_{m,m+1}}{\alpha}\alpha_{2q} & \dots & \alpha_{mn} - \frac{\alpha_{m,m+1}}{\alpha}\alpha_{2n} & d_m - \frac{\alpha_{m,m+1}}{\alpha}d_2 \\ 0 & -\frac{r_{m+1}}{\alpha} & 0 & \dots & 0 & 0 & \dots & r_q - \frac{r_{m+1}}{\alpha}\alpha_{2q} & \dots & r_n - \frac{r_{m+1}}{\alpha}\alpha_{2n} & -z_0 - \frac{r_{m+1}}{\alpha}d_2 \end{pmatrix},$$

is also in canonical form. This manipulation is called *pivoting* on the $(2, m+1)$-th element. The corresponding basic solution is feasible due to the assumption that $d_2/\alpha_{2,m+1}$ is the minimal ratio. The new relative cost coefficients are automatically updated and vanish for the variables of the new basis. Observe that the new relative cost coefficient for x_2 is

positive; there is no reason bringing it back into the basis (this would result in jumping back to (3.19)). The objective value changes from z_0 to $z_0 + r_{m+1}d_2/\alpha$ with $\alpha > 0$, $r_{m+1} < 0$ and $d_2 \geq 0$. If $d_2 = 0$, the rightmost column is unchanged and the move has been degenerate; otherwise the objective value strictly decreases.

To illustrate the use of the simplex tableau, we solve again the program (3.9). The initial simplex tableau A_c^b

$$\begin{pmatrix} x & y & s_1 & s_2 & s_3 & b \\ -3 & \boxed{1} & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 21 \\ 5 & 2 & 0 & 0 & 1 & 45 \\ -1 & -1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.20}$$

is already in canonical form, so the values in the last row are already the relative cost coefficients. We choose to enter y into the basis, so we pivot on the element marked by a square, $\alpha = 1$. The pivoting amounts to subtracting from the second, third and fourth rows respectively $2/\alpha$, $2/\alpha$, and $-1/\alpha$ times respectively the second row, then divide the second row by α. The variable to leave the basis is s_1 and as its current value is 0, we see immediately that the move will be degenerate. The next tableau is

$$\begin{pmatrix} -3 & 1 & 1 & 0 & 0 & 0 \\ \boxed{7} & 0 & -2 & 1 & 0 & 21 \\ 11 & 0 & -2 & 0 & 1 & 45 \\ -4 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

The move is degenerate; we stayed at the solution $(0, 0, 0, 21, 45)$. The next move is to enter x instead of s_2. The value of α is now 7, leading to the fractional values in the next tableau

$$\begin{pmatrix} 0 & 1 & 1/7 & 3/7 & 0 & 9 \\ 1 & 0 & -2/7 & 1/7 & 0 & 3 \\ 0 & 0 & \boxed{8/7} & -11/7 & 1 & 12 \\ 0 & 0 & -1/7 & 4/7 & 0 & 12 \end{pmatrix}.$$

The simplex algorithm jumps to the solution $(3, 9, 0, 0, 12)$ with objective value -12. The next step leads to

$$\begin{pmatrix} 0 & 1 & 0 & 5/8 & -1/8 & 7.5 \\ 1 & 0 & 0 & -1/4 & 1/4 & 6 \\ 0 & 0 & 1 & -11/8 & 7/8 & 10.5 \\ 0 & 0 & 0 & 3/8 & 1/8 & 13.5 \end{pmatrix}.$$

This basic solution $(6, 7.5, 10.5, 0, 0)$ is optimal by Lemma 3.8. The relative cost coefficients are indeed the same as with the direct calculation (3.15).

The simplex algorithm proceeds as follows:

Step 0 Find a starting point i.e. a basic feasible solution x, and a basis B. Then using row manipulations transform the simplex tableau to its canonical form. (It is not always clear how to find a starting point. If all the constraints are given

by "$\leq$" inequalities then introducing slack variables will provide a starting point as in (3.20). Otherwise this can be a tricky point; a possible solution is by using **artificial variables** [9, Section 3.5]).

Step 1 Given the current basic solution x and the basis B, the relative cost coefficients are given by the bottom row of the simplex tableau. If they are all nonnegative, stop; the current solution is optimal by Lemma 3.8.

Step 2 Choose a variable j with $r_j < 0$. This nonbasic variable will become basic in the next step.

Step 3 Let $(\alpha_1, \ldots, \alpha_m)$ denote the j-th column of the tableau and $(d_1, \ldots, d_m)$ denote the last column (without the bottom row of the tableau). d_i is the value of the i-th basic variable. If $\alpha_i \leq 0$ for any i, stop; the problem is unbounded. Otherwise, calculate the ratios $\{d_i/\alpha_i : \alpha_i > 0\}$, choose i for which the ratio is minimal and let $\varepsilon = d_i/\alpha_i$. Then pivot on the (ij)-th element to obtain the new simplex tableau. x_j enters the basis instead of the variable whose column is the i-th unit vector. The objective function changes by $\varepsilon r_j \leq 0$, with equality holding if and only if $\varepsilon = 0$ if and only if the move is degenerate. Update x and B and return to Step 1.

In Step 3, if $\varepsilon = 0$ then x remains unchanged; the basis B, however, always changes. This can only occur when x is degenerate. If $\varepsilon > 0$ and i is not unique, then the new solution is degenerate; a different choice of i does not impact x but leads to a different basis B. Usually one chooses the minimal such i.
In Step 2 it is not specified which variable should enter the basis when more than one relative cost coefficient is negative. A common practice is to choose the one with the smallest value of r_j. The intuitive justification is that this is the variable we wish the most to be in the basis.

Now we can provide a complete proof that the simplex algorithm works under the assumption of non degeneracy.

Theorem 3.9 *Let* $\min c^t x$ *subject to* $Ax = b$ *and* $x \geq 0$ *be a linear program and* x *a basic solution of it. Assume that any basic feasible solution is non degenerate and has exactly* m *nonzero coordinates. Then the simplex algorithm terminates after a finite number of steps either at an optimal solution or by determining unboundedness of the problem.*

Proof. If the algorithm does not stop, then the non degeneracy assumption gives $\varepsilon > 0$ in Step 3 hence a strict decrease of the objective function. This means that the algorithm never revisits a solution it has visited before. Since there are finitely many basic solutions, this can only happen finitely many times, after which the algorithm stops. But the algorithm stops only when unboundedness of the problem is determined or when an optimal solution is reached.

In case of degeneracy, an optimal solution can fail to satisfy the optimality criterion of Lemma 3.8. We provide an example to illustrate this situation, as well as the corresponding steps of the simplex method. Consider the program

$$\min x_1 - x_2 + x_3 + 4x_4 \qquad \text{subject to} \quad x \geq 0 \quad \text{and} \quad \begin{pmatrix} 1 & 0 & 3 & 1 \\ 0 & 1 & 1 & -1 \end{pmatrix} x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The program is bounded because the first constraint implies $x_1, x_3, x_4 \leq 1$ and the second implies $x_2 \leq x_4$. Therefore the optimal value is attained at a basic feasible solution. The basic solutions are obtained by letting two of the variables vanish and solving for the rest, yielding

$(1,0,0,0)$	with value	1
$(0,0,1/4,1/4)$	with value	$5/4$
$(0,-1/3,1/3,0)$	infeasible	
$(0,1,0,1)$	with value	3.

One sees that $(1,0,0,0)$ is the optimal solution to the problem. We present the steps of the simplex algorithm for this example. Clearly the first two columns are a basis of $\mathbb{R}^2$, corresponding to the degenerate solution $(1,0,0,0)$. The canonical form of the simplex tableau is obtained from the initial tableau by adding the second row and subtracting the first row from the third

$$\begin{pmatrix} 1 & 0 & 3 & 1 & 1 \\ 0 & 1 & 1 & -1 & 0 \\ 1 & -1 & 1 & 4 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 & 0 & 3 & 1 & 1 \\ 0 & 1 & \boxed{1} & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \end{pmatrix}. \tag{3.21}$$

Pivoting on the element marked by the square leads to a degenerate move and the determination of unique optimality by Lemma 3.8:

$$\begin{pmatrix} 1 & -3 & 0 & 4 & 1 \\ 0 & 1 & 1 & -1 & 0 \\ 0 & 1 & 0 & 1 & -1 \end{pmatrix}.$$

This example shows that the uniqueness is with respect to the basic solution and not to the basis. The relative cost coefficients with respect to the degenerate solution $(1,0,0,0)$ are all positive if we let x_3 be basic; if x_2 is basic then one relative cost coefficient is negative and the other is positive.

3.7 Degeneracy

As we have seen, the simplex algorithm works well if no degenerate solutions are present. The existence of degenerate solutions may cause it to fail; it is possible that the algorithm will jump from one degenerate solution to another and back thus entering a loop (**cycling**). When describing the algorithm, it was not specified which variable should enter the basis if more than one relative cost coefficient is strictly negative. Furthermore, in case of ties

as for the index i that minimizes the ratios d_i/α_i, the choice of the variable to leave the basis is arbitrary, too.

As mentioned above, the common practice is to choose the variable with the smallest relative cost coefficient and, in case of ties, to choose the minimal index i among the tied variables. In practice, using these decision rules the algorithm usually does not cycle, even when degenerate solutions exist. However, one can find examples when cycling does occur with these rules applied [15, Example 4.6].

Bland [3] proposed a simple rule that prevents cycling: when several r_j's are negative, choose the minimal j, and do the same in case of ties for d_i/α_i. Another procedure that deals with cycling involves slight perturbations of the vector b in a way that the solution for the perturbed system is non degenerate [9, Chapter 3, Exercises 15–17].

The reader is referred to these references for proofs and more details.

3.8 Worst case scenario for the simplex method

We conclude this section with a discussion about the runtime of the simplex method. Each step of the algorithm consists of elementary arithmetic operations on an $(m+1)\times(n+1)$ matrix and the number of such operations is clearly bounded by a polynomial in m and n. We therefore concentrate the discussion on the number of steps required until an optimal solution is reached.

In practice, in most cases the number of basic solutions visited by the algorithm is bounded by a linear function of m. However, in some cases the algorithm may need exponentially many steps before arriving at the optimal solution. In this subsection we present an example where this is indeed the case. We consider a family of linear programs with $2n$ variables and n constraints, having 2^n basic solutions and such that the simplex algorithm runs through all of them until the optimal solution is found. The example is taken from [9, Chapter 5.2].

We begin by presenting the program for the particular case $n = 3$:

$$\min -100x_1 - 10x_2 - x_3 \qquad \text{subject to} \quad x \geq 0 \quad \text{and}$$

$$\begin{array}{rcrcrl} x_1 & & & & & \leq 1 \\ 20x_1 & + & x_2 & & & \leq 100 \\ 200x_1 & + & 20x_2 & + & x_3 & \leq 10000. \end{array}$$

Introducing slack variables $s = (s_1, s_2, s_3)$, a basic feasible solution is given by $x = 0 \in \mathbb{R}^3$ and $s = (1, 100, 10000)$. The corresponding tableau is

$$\begin{pmatrix} \boxed{1} & 0 & 0 & 1 & 0 & 0 & 1 \\ 20 & 1 & 0 & 0 & 1 & 0 & 100 \\ 200 & 20 & 1 & 0 & 0 & 1 & 10000 \\ -100 & -10 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

A moment's thought is sufficient to conclude that s_3 should be zero in any optimal solution, and another moment's thought might be needed to see that $x = (0, 0, 10000)$ and $s = 0$ is the optimal solution. However, the simplex algorithm applied to this program (selecting the variable with smallest r_j at each step) pivots on the element marked with a square. The resulting sequence of tableaus follows with the basic variables in the top row.

$$\begin{pmatrix} x_1 & & & & s_2 & s_3 & \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & \boxed{1} & 0 & -20 & 1 & 0 & 80 \\ 0 & 20 & 1 & -200 & 0 & 1 & 9800 \\ 0 & -10 & -1 & 100 & 0 & 0 & 100 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 & x_2 & & & & s_3 & \\ 1 & 0 & 0 & \boxed{1} & 0 & 0 & 1 \\ 0 & 1 & 0 & -20 & 1 & 0 & 80 \\ 0 & 0 & 1 & 200 & -20 & 1 & 8200 \\ 0 & 0 & -1 & -100 & 10 & 0 & 900 \end{pmatrix} \Longrightarrow$$

$$\begin{pmatrix} & x_2 & & s_1 & & s_3 & \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 20 & 1 & 0 & 0 & 1 & 0 & 100 \\ -200 & 0 & \boxed{1} & 0 & -20 & 1 & 8000 \\ 100 & 0 & -1 & 0 & 10 & 0 & 1000 \end{pmatrix} \Longrightarrow \begin{pmatrix} & x_2 & x_3 & s_1 & & & \\ \boxed{1} & 0 & 0 & 1 & 0 & 0 & 1 \\ 20 & 1 & 0 & 0 & 1 & 0 & 100 \\ -200 & 0 & 1 & 0 & -20 & 1 & 8000 \\ -100 & 0 & 0 & 0 & -10 & 1 & 9000 \end{pmatrix} \Longrightarrow$$

$$\begin{pmatrix} x_1 & x_2 & x_3 & & & & \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -20 & \boxed{1} & 0 & 80 \\ 0 & 0 & 1 & 200 & -20 & 1 & 8200 \\ 0 & 0 & 0 & 100 & -10 & 1 & 9100 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 & & x_3 & & s_2 & & \\ 1 & 0 & 0 & \boxed{1} & 0 & 0 & 1 \\ 0 & 1 & 0 & -20 & 1 & 0 & 80 \\ 0 & 20 & 1 & -200 & 0 & 1 & 9800 \\ 0 & 10 & 0 & -100 & 0 & 1 & 9900 \end{pmatrix} \Longrightarrow$$

$$\begin{pmatrix} & & x_3 & s_1 & s_2 & & \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 20 & 1 & 0 & 0 & 1 & 0 & 100 \\ 200 & 20 & 1 & 0 & 0 & 1 & 10000 \\ 100 & 10 & 0 & 0 & 0 & 1 & 10000 \end{pmatrix}.$$

The solution $x = (0, 0, 10000)$ and $s = (1, 100, 0)$ with objective value -10000 is (uniquely) optimal by Lemma 3.8.

The algorithm visited eight basic solutions until optimality is achieved. In fact, it is not difficult to see that these are all the basic feasible solutions of the program: for each j either x_j or s_j (but not both) should be in the basis. To see this, observe firstly that $x_1 + s_1 = 1$, so that at least one of them has to be present in the basis. Furthermore, if $x_2 = s_2 = 0$ then the constraint $20x_1 + x_2 + s_2 = 100$ means that $x_1 = 5$, an infeasible solution. Therefore one of x_2 and s_2 has to be present in the basis as well. Similarly, if $x_3 = s_3 = 0$ then $200x_1 + 20x_2 = 10000$, or $10x_1 + x_2 = 500$, which is inconsistent with $x_2 \leq 100$ and $x_1 \leq 1$. It follows that one of x_3 and s_3 has to be present in the basis, and since the basis consists of exactly three elements this assertion follows.

For general n, the program is

$$\min -10^{n-1}x_1 - 10^{n-2}x_2 - \cdots - 10x_{n-1} - x_n \quad \text{subject to} \quad x \geq 0 \quad \text{and}$$

$$\begin{array}{rcrcrcrcl}
x_1 & & & & & & & \le & 1 \\
20x_1 & + & x_2 & & & & & \le & 100 \\
200x_1 & + & 20x_2 & + & x_3 & & & \le & 10000 \\
\vdots & & \vdots & & & & & \vdots & \\
2\Big(10^{i-1}x_1 & + & 10^{i-2}x_2 & + & \ldots & + & 10x_{i-1}\Big) & + \quad x_i & \le 100^{i-1}, \quad i \le n.
\end{array}$$

Or, more compactly,

$$\begin{aligned}
&\min -\sum_{i=1}^{n} 10^{n-i}x_i \qquad \text{subject to} \quad x \ge 0 \quad \text{and} \\
&2\sum_{j=1}^{i-1} 10^{i-j}x_j + x_i \le 100^{i-1}, \qquad i \in [n] \stackrel{\text{def}}{=} \{1,\ldots,n\}.
\end{aligned} \tag{3.22}$$

Lemma 3.10 *The linear program (3.22) has exactly 2^n basic feasible solutions.*

Proof. The proof follows exactly the same lines as for the case $n = 3$; we show that for any i at least one of the variables x_i and s_i has to be present in the basis. Firstly, observe that the constraints imply that $x_j \le 100^{j-1} = 10^{2j-2}$ for $j = 1,\ldots,n$. If $x_i = s_i = 0$ then the i-th constraint is satisfied as equality, so

$$10^{2i-2} = 100^{i-1} = 2\sum_{j=1}^{i-1} 10^{i-j}x_j \le 2\sum_{j=1}^{i-1} 10^{i-j}10^{2j-2} = 2\sum_{j=1}^{i-1} 10^{i+j-2} < 10^{2i-2}, \tag{3.23}$$

since $2\sum_{k=1}^{t} 10^k < 10^{t+1}$ for any integer t. It follows from the above contradiction that for any i either x_i or s_i should be in the basis, and they cannot both be there because a basis consists of n elements only. Therefore there are at most 2^n basic solutions to this program.
Now observe that if $s_1 = 0$ then $x_1 = 1 > 0$, and if $s_i = 0$ for some $i > 1$, then the i-th constraint, holding as equality, is

$$x_i = 100^{i-1} - 2\sum_{j=1}^{i-1} 10^{i-j}x_j > 0$$

by (3.23). It follows that for any $J \subseteq [n]$ the selection $s_J = x_{[n]\setminus J} = 0$ leads to a feasible non degenerate basic solution, and the lemma follows.

We leave it to the reader to verify that

Proposition 3.11 *Apply the simplex method to the linear program (3.22) with initial basic solution $x = 0$ and $s_j = 100^{j-1}$. At each step let the variable with the smallest (i.e. most negative) relative cost coefficient enter the basis. Then the algorithm visits all 2^n vertices before terminating at the optimal solution given by*

$$x = (0, 0, \ldots, 100^{n-1}) \qquad s = (1, 100, \ldots, 100^{n-2}, 0).$$

The reason why the simplex method is so inefficient in this example is this: because of the structure of the cost vector c, it is ten times more advantageous to have x_i basic rather than to have x_{i+1}. The constraints, however, allow x_{i+1} to be a hundred times larger than x_i, which greatly exceeds the benefit from the latter due to the cost vector. The algorithm keeps trying to enter variables with lower indices until it "has no choice", which requires this large number of steps.

The example above illustrates that in general one cannot count on the simplex algorithm to give efficient solutions. In contrast with this somewhat pathological example, in the next section we show that the discrete Monge–Kantorovich problem can be solved efficiently. Note that in the Monge–Kantorovich problem the cost function is arbitrary; the algorithm presented in the next section takes advantage only of the special form of the constraints matrix.

4 The Hungarian method

The special structure of the discrete Monge–Kantorovich problem as a linear program gives rise to other, more efficient methods than the simplex method. Since a permutation σ can be seen as an assignment of i to $\sigma(i)$, this problem is often referred to in the literature as **the assignment problem**; in our context, assigning the point x_i to the point y_j incurs a cost of c_{ij}. In this section, we present the algorithm outlined by Munkres [11], which is a variant of the one given by Kuhn [8]. Kuhn named his algorithm "The Hungarian method" because his algorithm was inspired by the work of two Hungarian mathematicians, Kőnig and Egerváry, that preceded linear programming by more than a decade. We recall that the discrete Monge–Kantorovich is to find a **bistochastic** matrix, i.e. a matrix $(x_{ij})_{i,j=1}^n$ with

$$x_{ij} \geq 0, \qquad \forall i \sum_{j=1}^{n} x_{ij} = 1 \quad \text{and} \quad \forall j \sum_{i=1}^{n} x_{ij} = 1 \tag{4.1}$$

and such that $\sum_{i,j=1}^n c_{ij}x_{ij}$ is minimal. $C = (c_{ij})_{i,j=1}^n$ is the cost matrix.
The first important observation is that one can add or subtract a constant from any row or column of the cost matrix without changing the optimal solution.

Lemma 4.1 *Let $C = (c_{ij})$ be the cost matrix. Suppose that we subtract p_i from the i-th row of C and q_j from the j-th column, where p_i and q_j are arbitrary numbers. In other words, we replace c_{ij} by $\tilde{c}_{ij} = c_{ij} - p_i - q_j$. Then the optimal solutions remain the same.*

Proof. Let (x_{ij}) satisfy (4.1), then

$$\sum_{i,j=1}^{n} \tilde{c}_{ij}x_{ij} = \sum_{i,j=1}^{n} c_{ij}x_{ij} - \sum_{i=1}^{n} p_i \sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} q_j \sum_{i=1}^{n} x_{ij} =$$

$$= \sum_{i,j=1}^{n} c_{ij}x_{ij} - P - Q, \qquad \text{where} \qquad P = \sum_{i=1}^{n} p_i \quad \text{and} \quad Q = \sum_{j=1}^{n} q_j.$$

Thus minimizing with respect to $\tilde{C}$ is equivalent to minimizing with respect to C.

If the entries of $\tilde{C}$ are nonnegative and there exists a permutation $\sigma \in S_n$ such that $\tilde{c}_{i,\sigma(i)} = 0$ for all i, then σ is optimal. We refer to σ as a *zero-permutation*. Our goal is to find p_i's and q_j's so that the resulting matrix, with values $\tilde{c}_{ij} = c_{ij} - p_i - q_j$, be nonnegative and admit a zero-permutation σ. σ will then be optimal, and its cost with respect to C is $P + Q = \sum_{i=1}^n p_i + \sum_{j=1}^n q_j$.
Recall that the Kantorovich duality says that

$$\inf_{\sigma \in S_n} \sum_{i=1}^{n} c_{i,\sigma(i)} = \sup \left\{ \sum_{i=1}^{n} \varphi(i) + \sum_{j=1}^{n} \psi(j) \right\},$$

where the supremum is on the set of functions $\varphi, \psi : \{1, \dots, n\} \to \mathbb{R}$ satisfying

$$\forall i, j \qquad \varphi(i) + \psi(j) \leq c_{ij}.$$

Therefore, the numbers p_i and q_j, $i, j = 1, \dots, n$ are precisely the variables of the Kantorovich dual problem.

A first step for finding such values of p_i and q_j is this: for each i let p_i be the minimal entry in row i. Then the matrix $\tilde{C}$ defined by $c_{ij} - p_i$ (i.e. the result of subtracting p_i from row i of C) is nonnegative and has a zero in each of its rows. Now for each j let q_j be the minimal entry in column j of $\tilde{C}$, and set $\tilde{C} = c_{ij} - p_i - q_j$ (i.e. subtract from each column of $\tilde{C}$ its minimal element). The resulting $\tilde{C}$ now has zeros in every row and in every column. For 2×2 matrices this guarantees the existence of a zero-permutation, but this fails to hold in general for matrices of larger order. For example, consider the following matrix to which the above procedure is applied:

$$C = \begin{pmatrix} 2 & 5 & 4 \\ 6 & 11 & 3 \\ 7 & 9 & 4 \end{pmatrix} \Rightarrow p = (2, 3, 4) \Rightarrow \tilde{C}_1 = \begin{pmatrix} 0 & 3 & 2 \\ 3 & 8 & 0 \\ 3 & 5 & 0 \end{pmatrix} \Rightarrow q = (0, 3, 0) \Rightarrow \tilde{C}_2 = \begin{pmatrix} 0 & 0 & 2 \\ 3 & 5 & 0 \\ 3 & 2 & 0 \end{pmatrix}.$$

Unfortunately, $\tilde{C}_2$ has no zero-permutations despite having zeros in any of its rows and any of its columns. We can "fix" $\tilde{C}_2$ to have a zero-permutation by decreasing p_1 by 2 and increasing q_1 and q_2 by 2, yielding

$$p = (0, 3, 4), \quad q = (2, 5, 0), \qquad \tilde{C} = \begin{pmatrix} 0^* & 0 & 4 \\ 1 & 3 & 0^* \\ 1 & 0^* & 0 \end{pmatrix},$$

for which (23) is a zero-permutation (the corresponding zeros are "starred"). The reader can easily check that this permutation incurs the minimal cost, $2 + 9 + 3 = 14$, with respect to the original matrix C. Indeed $14 = \sum_{i=1}^{3} p_i + \sum_{j=1}^{3} q_j$.

Before presenting Munkres' algorithm we need to introduce some definitions. In the following, a **line** is a horizontal or vertical line in the matrix, i.e. a row or a column. For a matrix $C \in M_n(\mathbb{R})$ with nonnegative entries, a collection of zeros is said to be **independent** if any pair them does not lie in the same line. A **cover** of C is a collection of lines such that every zero entry in the matrix is covered by a line. A minimal cover of C is a cover of minimal size, i.e. one containing as few lines as possible. Similarly, a maximal independent set is an independent set of maximal size. Note carefully that these definitions are with respect to *size*, not set inclusion. As an example, consider the following matrix and its two covers

$$C = \begin{matrix} 0^* & 0 & \tilde{0} \\ \tilde{0} & 7 & 1 \\ 0 & 2 & 6 \end{matrix}, \qquad \begin{matrix} \not{0} & \not{0} & \not{0} \\ \not{0} & 7 & 1 \\ \not{0} & 2 & 6 \end{matrix}, \qquad \begin{matrix} \not{0} & \not{0} & \not{0} \\ 0 & 7 & 1 \\ 0 & 2 & 6 \end{matrix}.$$

The starred zero forms an inclusion-wise maximal independent set; none of the other zeros in the matrix can be added to it and maintain independence. However, the set of $\tilde{0}$'s is larger and independent. Hence the starred zero is *not* a maximal independent set according to our definition. Similarly, the left cover is inclusion-wise minimal but not minimal, since the cover to the right has a smaller size.
It is clear that no line can cover more than one independent zero; hence if an independent set of size k exists then any cover contains at least k lines. The nontrivial converse also holds:

Theorem 4.2 (Kőnig) *For $C \in M_n(\mathbb{R})$, the size of a minimal cover of C is equal to the size of a maximal independent set in C.*

It is more convenient to prove Theorem 4.2 in terms of bipartite graphs. The reader not familiar with graph theory need not worry; all the notions are introduced in Definition 4.3, and no result from graph theory needs to be invoked. In fact, the algorithm does not explicitly use Theorem 4.2 but is rather inspired from it. The reader can safely skip this part and jump directly to the algorithm; no completeness will be lost.

Definition 4.3 *A (finite) graph is an ordered pair $G = (V, E)$ where V is a finite set of **vertices** and E is a collection of unordered pairs from V. The elements of E are called **edges**. If $\{u, v\} \in E$, then u is **adjacent** to v, and vice versa (we do not allow any vertex to be adjacent to itself). A **path** from u to v in G is a finite sequence*

$$(u = u_0, u_1, \ldots, u_k = v)$$

*where u_{i-1} is adjacent to u_i. When such a path exists, v is **reachable** from u, and vice versa; this means that starting from u, one can travel along edges of the graph to arrive at v. A path is **simple** if all of its vertices $\{u_i\}_{i=0}^{k-1}$ are distinct (we allow in general simple nontrivial paths from u to itself, but such paths will not be needed). It is clear that any path contains a simple path. For $U \subseteq V$, $v \in V$ is reachable from U if it is reachable from some $u \in U$.*
*A graph is **bipartite** if V can be written as a nontrivial disjoint union $S \cup T$ such that no two elements from S and no two elements from T are adjacent; edges can only exist between elements of S and elements of T. For example, one can think of S as a set of workers and T as a set of jobs, S_i being connected to T_j if and only if worker i is qualified for job j (this is Kuhn's motivation in [8]). We denote such graphs by $G = (S, T; E)$.*
*A **directed** graph is a graph whose edges have orientations (formally, E is then a collection of ordered pairs from V). This allows breaking the symmetry of the edges; in a directed graph u can be adjacent to v without v being adjacent to u. The notions of "path" and "reachable" also extend accordingly, taking into account the orientation of the edges.*
*A **matching** in a graph is a collection of edges $M \subseteq E$ such that no two contain a common vertex. For vertices covered by the matching we abuse notation and write $v \in M$.*

Finally, a ***vertex cover*** *is a subset $U \subseteq V$ of vertices covering all the edges, in the sense that for any $e \in E$ there exists $u \in U$ such that $u \in e$.*

The relation between the above notions and covers and matrices is rather straightforward. Given a matrix $C \in M_n(\mathbb{R})$, form a bipartite graph $G = (S, T; E)$ as follows. Let $S = \{s_1, \ldots, s_n\}$ represent the columns of C and $T = \{t_1, \ldots, t_n\}$ represent its rows. The edges of G are identified with the zeros of the matrix C; s_i is adjacent to t_j if and only if $c_{ij} = 0$. The resulting graph is bipartite; no two columns and no two rows are adjacent. If two edges share a common vertex, t_j, say, then the two corresponding zeros lie both in row j; they cannot be independent. Therefore, independent sets in the matrix are equivalent to matchings in G. A path from column i to row j consists of alternating horizontal and vertical jumps between zeros in the matrix. For example,

$$\begin{pmatrix} * & * & 0_5 & * & 0_6 \\ 0_1 & * & * & 0_2 & a \\ * & * & * & * & 0_7 \\ * & * & * & * & * \\ * & * & 0_4 & 0_3 & * \end{pmatrix} \tag{4.2}$$

is a path from column 1 to row 3 requiring seven steps. The first zero, denoted by 0_1, is the edge between column 1 and row 2, and so on. The path is

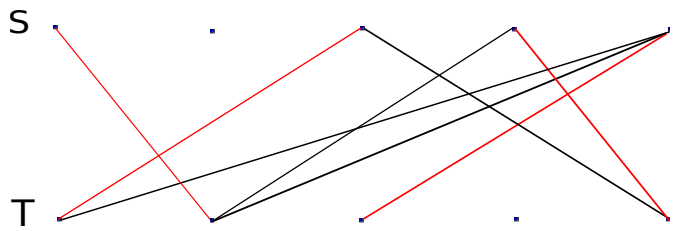
$$\text{column } 1 \Rightarrow \text{row } 2 \Rightarrow \text{column } 4 \Rightarrow \text{row } 5 \Rightarrow \text{column } 3 \Rightarrow \text{row } 1 \Rightarrow \text{column } 5 \Rightarrow \text{row } 3.$$

If $a = 0$ then another path exists through a, requiring three steps only:

$$\text{column } 1 \Rightarrow \text{row } 2 \Rightarrow \text{column } 5 \Rightarrow \text{row } 3.$$

Paths from columns to columns, rows to rows and rows to columns are analogous. The bipartite graph corresponding to the matrix (4.2) when $a = 0$ is presented in Figure 2. The red edges form a matching. They correspond to the independent zeros 0_1, 0_3, 0_5 and 0_7 in (4.2).

Figure 2: Graph representation of the matrix (4.2). S represents the column and T the row



Finally, a vertex cover of the bipartite graph is equivalent to a cover with lines as defined above.

We are now ready to state the equivalent of Theorem 4.2 in terms of graph theory.

Theorem 4.4 *For a bipartite graph $G = (S, T; E)$, the maximal size of a matching equals the minimal size of a vertex cover.*

Remark 6 *It is not assumed that S and T have the same size. Indeed, Theorem 4.2 holds for arbitrary rectangular matrices.*

Proof. A matching M is a set of edges with no common vertex, so no vertex can cover more than one of the edges in M. It follows that the size of any vertex cover is not less than the size of any matching.
To prove the theorem, we present a constructive proof for the existence of a matching M and a covering subset U of the same size.
At any stage of the algorithm we have a matching M associated with a *directed* bipartite graph G_M. G_M has the same edges as G, oriented as follows: any edge in M is directed from T to S, while all the others are directed from S to T. Throughout this proof, when we name a vertex s (respectively t) we assume implicitly that it is a member of S (respectively T).

Step 1 Initialize $M = \phi$; all the edges of G_M are oriented from S to T.

Step 2 For a matching M and the corresponding G_M we let $R = R_M$ be the set of vertices that are not involved in M, $R_S = R \cap S$ and $R_T = R \cap T$. We consider the set $Z = Z_M$ of vertices that are reachable from R_S in the directed graph G_M

$$Z \stackrel{\text{def}}{=} \{v \in S \cup T : \exists s \in R_S \text{ and a path from } s \text{ to } v \text{ in } G_M\},$$

and consider the intersection $Z \cap R_T$.

If $Z \cap R_T$ is empty, go to Step 3.

If $Z \cap R_T \neq \phi$, let $(s = v_0, v_1, \ldots, v_k = t)$ be a path where $s \in R_S$ and $t \in R_T$. As G is bipartite, $v_{2i} \in S$ and $v_{2i+1} \in T$ for any i, so that k is odd. Since any path contains a simple path, we may assume without loss of generality that the path above is simple. By definition of G_M, the only way to get from T to S is by using edges from M, so the edge $\{v_{2i+1}, v_{2i+2}\} \in M$ for any i. Form another matching M' as the symmetric difference between M and the path: add to M any edge of the form $\{v_{2i}, v_{2i+1}\}$ and remove any edge of the form $\{v_{2i+1}, v_{2i+2}\}$. As k is odd M' has precisely one more element than M.

To see that M' is a matching, observe firstly that the collection $\{\{v_{2i}, v_{2i+1}\}\}_{i=0}^{(k-1)/2}\}$ is a matching, since the path is simple. Furthermore, all the vertices in the path do not appear in any of the other edges in M': this is true for v_0 and v_k by the hypothesis that $s \in R_S$ and $t \in R_T$. The other vertices $v_1, \ldots, v_{k-1}$ only appear in M in edges of the form $\{v_{2i+1}, v_{2i+2}\}$, which are not members of M'. This means that M' is a matching, and it has been shown in the last paragraph that it has one

more element than M. Replace M by M', update G_M accordingly and run again Step 2. Since the size of any matching cannot exceed the size of S, Step 2 can only be repeated finitely many times.

Step 3 At this point $Z \cap R_T = \phi$, and we claim that the set $U = (T \cap Z) \cup (S \setminus Z)$ is (1) a vertex cover of G and (2) has the same size as M.

(1) If $s \in Z$ and $\{s,t\} \in E$ then t too is in Z: if $\{s,t\} \in M$ then s can only be reached through t; otherwise t is reachable through s. This means that $T \cap Z$ covers all the edges with vertices in $S \cap Z$. Since $S \setminus Z$ covers all other edges, it follows that U is a vertex cover (we have not used $Z \cap R_T = \phi$ here).

(2) In G_M, it is only possible to reach a vertex $s \notin R_S$ through a unique vertex $t \notin R_T$. Conversely, once $t \notin R_T$ is reached, the unique vertex $s \notin R_S$ satisfying $\{s,t\} \in M$ is reachable, too. Thus

$$|Z \cap S \setminus R_S| = |Z \cap T \setminus R_T|.$$

Since $R_S \subseteq Z$, it follows that

$$|M| = |S \setminus R_S| = |S \cap Z \setminus R_S| + |S \setminus (Z \cup R_S)| = |Z \cap T \setminus R_T| + |S \setminus Z|,$$

which in fact holds for an arbitrary matching M. Since $Z \cap R_T = \phi$ the above equality simplifies to $|M| = |Z \cap T| + |S \setminus Z| = |U|$ as desired.

In order to make the proof truly constructive, we must describe the procedure of determining the set Z. We will do this using the terminology of matrices, zeros and covers. Munkres' algorithm for finding a maximal independent zero set and a minimal cover follows. As stated above, the proof of its correctness does not use Theorem 4.2.

Description of the algorithm

The input is a nonnegative real $n \times n$ matrix C ($n > 1$), and at any step some of the following objects are manipulated:

- An integer index r.
- A real number p_i for every row i and a real number q_j for every column j.
- An independent zero set M.
- A collection of lines U (*covered lines*).

The following statements hold throughout the runtime of the algorithm:

- For any i and any j, $p_i + q_j \leq c_{ij}$.
- U and M have the same size.
- Every independent zero in M is on a covered line.

U can be seen as a partial cover, since not necessarily all the zeros in the matrix are covered. Once U is a cover, we have by (the trivial part of!) Theorem 4.2 that M is maximal and U is minimal. Following [11], we refer to the elements of M as "starred zeros". Integer values will be assigned to some of the zeros (but never to the elements of M); we refer to these zeros as *labeled*.

The algorithm consists of four steps. When we write "the matrix", we refer to the matrix $\tilde{C}$ with elements $\tilde{c}_{ij} = c_{ij} - p_i - q_j$, a nonnegative matrix by the hypothesis on p_i and q_j. The first step is the initialization; once done, the algorithm never returns to it. In the second step some vertical lines are replaced by horizontal ones, and some zeros become labeled. In the third step all horizontal lines are removed and replaced by vertical ones, and M increases its size by one. All labels are removed as well. In the fourth step M and U do no change; only the matrix $\tilde{C}$ does (through p_i and q_j). Steps 2 and 3 correspond to Kőnig's theorem. Step 4 is inspired by Egerváry's work.

Step 1 Let p_i be the minimal element of the i-th row of C, $q_j = 0$ for any j, and update $\tilde{C}$ by $\tilde{c}_{ij} = c_{ij} - p_i$. Now let q_j be the minimal element of the j-th column of $\tilde{C}$, and update again $\tilde{c}_{ij} = c_{ij} - p_i - q_j$. The requirements from p_i and q_j are satisfied, and $\tilde{C}$ now has at least one zero in each row and in each column. Choose an independent set M column-wise; that is, scan the columns by increasing order and look for zeros. Once one is found that is independent from all others, star it. Let M be the independent set of the starred zeros, and cover its corresponding columns to form U. Set the index r to equal 1; no zeros are labeled.

Step 2 Given an independent set M and a cover U, look for an uncovered zero e. If there is none, M is maximal and U is minimal. If the size of M is n (the order of the matrix), stop; the required zero-permutation is given by M. If the size of M is less than n, the matrix admits no zero-permutation and the values of p_i and q_j need to be changed; go to Step 4.

If an uncovered zero e is found, label it by the index r, increase r by one and look for a starred zero in the row of e. If there is none, we can find an independent set larger than M; go to Step 3. If there is a starred zero Z^* that shares a row with e, the column of Z^* must be covered (because Z^* is covered and e is not). Uncover the column of Z^* and cover instead its row, so now e is covered. Repeat this step with the updated U. Note that the only possible change in U at this step is replacing vertical lines with horizontal ones. The order by which the uncovered zeros are chosen is arbitrary.

Step 3 At this point we can construct a sequence of alternating labeled and starred zeros as follows. Let e_0 be the labeled zero due to which the algorithm arrived here from Step 2. Given a labeled zero e_{2j}, if there is no starred element in its column, stop the

sequence. Otherwise, the starred element has to be unique; call it e_{2j+1}. Now, when e_{2j} has been labeled it was uncovered, so the column of e_{2j+1} was uncovered. By the instructions in Step 2, this can only happen if some zero in the row of e_{2j+1} had been labeled before, thus replacing the vertical line covering e_{2j+1} with a horizontal one. This labeled zero, which we denote by e_{2j+2}, has to be unique, because after the labeling the row becomes covered. Furthermore the label of e_{2j+2} is strictly less than that of e_{2j}; in particular these zeros are distinct and hence the sequence is finite; it terminates at some e_{2k} with no starred zero in its column.

Let $A = \{e_{2j}\}$ and $B = \{e_{2j+1}\}$, then $B \subseteq M$ and $A \cap M = \phi$. By the preceding paragraph, A has one more element than B, and we claim that $M' = M \cup A \setminus B$ is an independent set. For brevity, we will prove this claim later. Star the elements of A and unstar those of B to replace M by M', and set U to be the columns of the elements of the new independent set. Remove all labels, and set $r = 1$. Return to Step 2 with the updated M and U.

Step 4 At this point M is maximal, U is a cover and the matrix needs to be modified because it admits no zero-permutations. Let $h > 0$ be the smallest uncovered element, subtract h from each uncovered column and add h to each covered row. This amounts to increasing q_j by h if column j is uncovered and decreasing p_i by h if row i is covered (equivalently, subtract h from any uncovered element and add h to every element that is covered twice). There are more uncovered columns than covered rows, because the size of the cover, say k, is strictly less than n; the sum $\sum p_i + \sum q_j$ increases by $h(n-k) > 0$. Thus $\tilde{C}$ can never change to a matrix it equalled previously during the runtime of the algorithm.

It is clear that the new $\tilde{C}$ is nonnegative, and has at least one uncovered zero. Furthermore, all starred zeros and all labeled zeros are covered once and hence remain zeros in $\tilde{C}$. Return to Step 2 with the modified matrix, without changing U, M and the labels.

Before proving the correctness and finiteness of the algorithm, we provide an example to illustrate it. The numbers above the arrows are the current value of $P + Q$, where $P = \sum_{i=1}^{n} p_i$ and $Q = \sum_{j=1}^{n} q_j$; once the algorithm terminates this value is the cost of the optimal permutation with respect to the original cost matrix C.
The first two iterations are the row and column subtraction and the initial M and U in Step 1. Then follow three iterations of Step 2.

$$\begin{array}{cccccccccccccccccc}
8 & 7 & 9 & 9 & & 1 & 0 & 2 & 2 & & 1 & 0^* & 2 & 0 & & \cancel{1} & \cancel{0^*} & \cancel{2} & \cancel{0_1} & & \cancel{1} & \cancel{0^*} & \cancel{2} & \cancel{0_1} & & \cancel{1} & \cancel{0^*} & \cancel{2} & \cancel{0_1} \\
5 & 2 & 7 & 8 & \overset{12}{\Rightarrow} & 3 & 0 & 5 & 6 & \overset{14}{\Rightarrow} & 3 & 0 & 5 & 4 & \overset{14}{\Rightarrow} & 3 & 0 & 5 & 4 & \overset{14}{\Rightarrow} & 3 & 0 & 5 & 4 & \overset{14}{\Rightarrow} & 3 & 0_3 & 5 & 4 \\
6 & 1 & 4 & 9 & & 5 & 0 & 3 & 8 & & 5 & 0 & 3 & 6 & & 5 & 0 & 3 & 6 & & 5 & 0 & 3 & 6 & & 5 & 0 & 3 & 6 \\
2 & 3 & 2 & 6 & & 0 & 1 & 0 & 4 & & 0^* & 1 & 0 & 2 & & 0^* & 1 & 0 & 2 & & \cancel{0^*} & \cancel{1} & \cancel{0_2} & \cancel{2} & & \cancel{0^*} & \cancel{1} & \cancel{0_2} & \cancel{2}
\end{array}$$

This last zero 0_3 has no starred zero in its row, so the algorithm moves to Step 3. The formed sequence is $e_0 = 0_3$, e_1 is the starred zero above e_0, and e_2 is 0_1. After carrying out Step 3, another iteration of Step 2 follows.

$$\begin{array}{c|ccc|c} & 1 & 0 & 2 & 0^* \\ \hline 3 & 0^* & 5 & 4 & \\ 5 & 0 & 3 & 6 & \\ 0^* & 1 & 0 & 2 & \end{array} \quad \begin{array}{c|ccc|c} & 1 & 0 & 2 & 0^* \\ \hline 3 & 0^* & 5 & 4 & \\ 5 & 0 & 3 & 6 & \\ \cancel{0^*} & \cancel{1} & \cancel{0_1} & \cancel{2} & \end{array}$$

Now all the zeros are covered and Step 4 is invoked. Here $h = 1$, q_1 and q_3 increase by 1 and p_4 decreases by 1, so the sum $P + Q$ increases by one. A new zero appears at the top left corner. After one iteration of Step 2, this new zero is labeled by the number 2. Then Step 4 needs to be invoked again, this time with $h = 2$; the sum $P + Q$ increases by two.

$$\begin{array}{c|ccc|c} & 0 & 0 & 1 & 0^* \\ \hline 2 & 0^* & 4 & 4 & \\ 4 & 0 & 2 & 6 & \\ \cancel{0^*} & \cancel{2} & \cancel{0_1} & \cancel{3} & \end{array} \quad \begin{array}{c|ccc|c} & \cancel{0_2} & \cancel{0} & \cancel{1} & \cancel{0^*} \\ \hline 2 & 0^* & 4 & 4 & \\ 4 & 0 & 2 & 6 & \\ \cancel{0^*} & \cancel{2} & \cancel{0_1} & \cancel{3} & \end{array} \quad \begin{array}{c|ccc|c} & \cancel{0_2} & \cancel{2} & \cancel{1} & \cancel{0^*} \\ \hline 0 & 0^* & 2 & 2 & \\ 2 & 0 & 0 & 4 & \\ \cancel{0^*} & \cancel{4} & \cancel{0_1} & \cancel{3} & \end{array} \quad \begin{array}{c|ccc|c} & \cancel{0_2} & \cancel{2} & \cancel{1} & \cancel{0^*} \\ \hline \cancel{0_3} & \cancel{0^*} & \cancel{2} & \cancel{2} & \\ 2 & 0 & 0 & 4 & \\ \cancel{0^*} & \cancel{4} & \cancel{0_1} & \cancel{3} & \end{array} \quad \begin{array}{c|ccc|c} & \cancel{0_2} & \cancel{2} & \cancel{1} & \cancel{0^*} \\ \hline \cancel{0_3} & \cancel{0^*} & \cancel{2} & \cancel{2} & \\ 2 & 0 & 0_4 & 4 & \\ \cancel{0^*} & \cancel{4} & \cancel{0_1} & \cancel{3} & \end{array}$$

Finally, Step 3 is invoked again to obtain a zero-permutation. The corresponding permutation is indicated on the original matrix too and has a cost of 17.

$$\begin{array}{c|ccc|c} & 0 & 2 & 1 & 0^* \\ \hline 0 & 0^* & 2 & 2 & \\ 2 & 0 & 0^* & 4 & \\ 0^* & 4 & 0_1 & 3 & \end{array} \qquad \begin{array}{cccc} 8 & 7 & 9 & 9^* \\ 5 & 2^* & 7 & 8 \\ 6 & 1 & 4^* & 9 \\ 2^* & 3 & 2 & 6 \end{array}$$

Remark 7 *The row and column reduction in Step 1 are not truly necessary. If C is nonnegative, one can use the simpler rule*

Step 1 *Set $r = 1$, $M = U = \phi$; no zeros are starred and there are no covering lines.*

This seems intuitively slightly less efficient, because the original Step 1 is an "easy" way to introduce many zeros to the matrix. If one uses the simpler rule then a possible sequence may be

$$\begin{array}{cccc} 8 & 7 & 9 & 9 \\ 5 & 2 & 7 & 8 \\ 6 & 1 & 4 & 9 \\ 2 & 3 & 2 & 6 \end{array} \quad \begin{array}{cccc} 7 & 6 & 8 & 8 \\ 4 & 1 & 6 & 7 \\ 5 & 0 & 3 & 8 \\ 1 & 2 & 1 & 5 \end{array} \quad \begin{array}{cccc} 7 & 6 & 8 & 8 \\ 4 & 1 & 6 & 7 \\ 5 & 0_1 & 3 & 8 \\ 1 & 2 & 1 & 5 \end{array} \quad \begin{array}{c|ccc} 7 & 6 & 8 & 8 \\ \hline 4 & 1 & 6 & 7 \\ 5 & 0^* & 3 & 8 \\ 1 & 2 & 1 & 5 \end{array} \quad \begin{array}{c|ccc} 6 & 6 & 7 & 7 \\ \hline 3 & 1 & 5 & 6 \\ 4 & 0^* & 2 & 7 \\ 0 & 2 & 0 & 4 \end{array}$$

$$\begin{array}{c|ccc} 6 & 6 & 7 & 7 \\ \hline 3 & 1 & 5 & 6 \\ 4 & 0^* & 2 & 7 \\ 0_1 & 2 & 0 & 4 \end{array} \quad \begin{array}{c|ccc} 6 & 6 & 7 & 7 \\ \hline 3 & 1 & 5 & 6 \\ 4 & 0^* & 2 & 7 \\ 0^* & 2 & 0 & 4 \end{array} \quad \begin{array}{c|ccc} 6 & 6 & 7 & 7 \\ \hline 3 & 1 & 5 & 6 \\ 4 & 0^* & 2 & 7 \\ \cancel{0^*} & \cancel{2} & \cancel{0_1} & \cancel{4} \end{array} \quad \begin{array}{c|ccc} 4 & 6 & 5 & 5 \\ \hline 1 & 1 & 3 & 4 \\ 2 & 0^* & 0 & 5 \\ \cancel{0^*} & \cancel{4} & \cancel{0_1} & \cancel{4} \end{array} \quad \begin{array}{c|ccc} 4 & 6 & 5 & 5 \\ \hline 1 & 1 & 3 & 4 \\ \cancel{2} & \cancel{0^*} & \cancel{0_2} & \cancel{5} \\ \cancel{0^*} & \cancel{4} & \cancel{0_1} & \cancel{4} \end{array}$$

$$\begin{array}{c|ccc} 3 & 5 & 4 & 4 \\ \hline 0 & 0 & 2 & 3 \\ \cancel{2} & \cancel{0^*} & \cancel{0_2} & \cancel{5} \\ \cancel{0^*} & \cancel{4} & \cancel{0_1} & \cancel{4} \end{array} \quad \begin{array}{c|ccc} 3 & 5 & 4 & 4 \\ \hline 0_3 & 0 & 2 & 3 \\ \cancel{2} & \cancel{0^*} & \cancel{0_2} & \cancel{5} \\ \cancel{0^*} & \cancel{4} & \cancel{0_1} & \cancel{4} \end{array} \quad \begin{array}{c|ccc} 3 & 5 & 4 & 4 \\ \hline 0^* & 0 & 2 & 3 \\ 2 & 0^* & 0 & 5 \\ 0 & 4 & 0^* & 4 \end{array} \quad \begin{array}{c|ccc} 3 & 5 & 4 & 1 \\ \hline 0^* & 0 & 2 & 0 \\ 2 & 0^* & 0 & 2 \\ 0 & 4 & 0^* & 1 \end{array} \quad \begin{array}{c|ccc} 3 & 5 & 4 & 1 \\ \hline \cancel{0^*} & \cancel{0} & \cancel{2} & \cancel{0_1} \\ 2 & 0^* & 0 & 2 \\ 0 & 4 & 0^* & 1 \end{array}$$

$$\begin{array}{cccc} 3 & 5 & 4 & 1 \\ 0^* & 0 & 2 & 0_1 \\ 2 & 0^* & 0 & 2 \\ 0_2 & 4 & 0^* & 1 \end{array} \overset{16}{\Rightarrow} \begin{array}{cccc} 3 & 5 & 4 & 1 \\ 0^* & 0 & 2 & 0_1 \\ 2 & 0^* & 0_3 & 2 \\ 0_2 & 4 & 0^* & 1 \end{array} \overset{16}{\Rightarrow} \begin{array}{cccc} 2 & 4 & 3 & 0 \\ 0^* & 0 & 2 & 0_1 \\ 2 & 0^* & 0_3 & 2 \\ 0_2 & 4 & 0^* & 1 \end{array} \overset{17}{\Rightarrow} \begin{array}{cccc} 2 & 4 & 3 & 0_4 \\ 0^* & 0 & 2 & 0_1 \\ 2 & 0^* & 0_3 & 2 \\ 0_2 & 4 & 0^* & 1 \end{array} \overset{17}{\Rightarrow}$$

leading finally to (another) optimal solution of cost 17

$$\overset{17}{\Rightarrow} \begin{array}{cccc} 2 & 4 & 3 & 0^* \\ 0^* & 0 & 2 & 0 \\ 2 & 0^* & 0 & 2 \\ 0 & 4 & 0^* & 1 \end{array} \qquad \begin{array}{cccc} 8 & 7 & 9 & 9^* \\ 5^* & 2 & 7 & 8 \\ 6 & 1^* & 4 & 9 \\ 2 & 3 & 2^* & 6 \end{array}$$

One sees that more steps were needed than with the original Step 1.

Before proceeding, we need to prove the missing part in Step 3.

Claim 1 *The set M' created in Step 3 is independent.*

Proof. Recall that $M' = (M \setminus B) \cup A$, where $A = \{e_{2j}\}$ are labeled zeros with decreasing labels and $B = \{e_{2j+1}\}$ are starred zeros. We firstly show that A is independent. By the instructions in Step 2, it is clear that no two labeled zeros can be in the same row. As the example above shows, two labeled zeros can share a column, but this cannot take place if they are both in A: for $e_{2l} \in A$ lying in column j, we show that no e_{2k}, $k > l$, can be in the same column j. This is clear if e_{2l} is the last element of the sequence, and if not, then the starred e_{2l+1} is in the column j, and e_{2l+2} is in e_{2l+1}'s row and a different column. Until e_{2l+2} has been labeled, column j was covered, so no e_{2k}, $k > l+1$ can be in column j. It follows that A is independent.
Since $M \setminus B$ is independent as a subset of M, we only have to show that $e \in A$ and $f \in M \setminus B$ cannot share a line. Suppose that the sequence terminates at e_{2k}. If $k = 0$ then $A = \{e_0\}$ is independent of M and the claim holds trivially. Otherwise, e_{2k} has a starred zero $e_{2k-1} \in B$ in its row and no starred zero in its column. Since no other starred zero can be in this row, e_{2k} is independent of $M \setminus B$. An analogous argument proves the same for e_0. For $0 < l < k$, e_{2l} has a starred zero, e_{2l+1}, in its column and another, e_{2l-1}, in its row. As both these zeros are in B, e_{2l} cannot share a line with any zero in $M \setminus B$.

It is clear that the algorithm terminates only with an independent set of size n and hence providing an optimal permutation. We proceed by showing its finiteness and polynomial runtime. Since the computational effort needed at any step is at most $O(n^2)$, it remains to show that each step is undertaken at most polynomially many times.
For Step 3 this is obvious; at each time we increase the size of M by one, and this size is bounded by n. Step 3 can thus be invoked at most n times.
As long as the algorithm does not go to Step 3, the size of M is constant. While this size is $k < n$, every iteration of Step 4 creates an uncovered zero so has to be followed by an iteration of Step 2. Then a vertical line is replaced by a horizontal one and as there

are k such lines, Step 2 can be iterated at most $k+1$ times. Thus, before moving from an independent set of size k to an independent set of size $k+1$ the algorithm does at most $2k+2$ steps. Summing up for $k = 0, \ldots, n-1$ we obtain the total number of steps done by the algorithm cannot exceed $n(n-1) = O(n^2)$. Therefore the total runtime is bounded by $O(n^2)O(n^2) = O(n^4)$.

Remark 8 *One may think that after Step 4 one should "remove all labels, set $r = 1$, uncover all rows and cover all columns of the independent zeros" (as at the end of Step 3) before returning to Step 2. This is however unnecessary. To see this, let U be the current covered lines and recall that any of the zeros that have been labeled in Step 2 is still a zero after the modification of Step 4. In Step 2 it is not specified how the zeros are chosen. If one "removes all labels, sets $r = 1$, uncovers all rows and covers all columns of the independent zeros" before returning to Step 2, it is possible that during Step 2 the zeros will be labeled in the exact same order and the resulting cover will be again U.*

5 Absolutely continuous measures on $\mathbb{R}^d$

Our goal in this section is to use the duality theorem in Section 2 in order to give a geometrical characterization of the optimal solutions for cost functions of the form $|x-y|^p$, $p>1$ (more generally, $c(|x-y|)$ where c is strictly convex). In particular, an important connection between the solution of the Monge–Kantorovich problem and its dual problem will be unveiled. Unsurprisingly, the case $p=2$ (*quadratic cost*) is of particular importance, and gives rise to a gradient descent algorithm due to Chartrand et al. [4] (see Section 6). Our approach here is, as in [14, Chapter 2], to develop the theory for quadratic cost and then sketch briefly the arguments for other convex costs. Another excellent reference is [6], where a different approach is used.

5.1 Quadratic cost function

As will be revealed immediately, the Monge–Kantorovich problem for quadratic cost is closely related to convexity of real valued functions on $\mathbb{R}^d$. For this reason, notions from convex analysis need to be introduced. We state all these results when they will be needed, and refer the reader to any standard convex analysis book, such as [12], for the proofs. For $x,y\in\mathbb{R}^d$ we use the compact notation xy to denote the standard inner product $\langle x,y\rangle=\sum_{i=1}^d x_iy_i$, and then $x^2=\langle x,x\rangle=\|x\|^2$. For reasons that will become clear very soon, we consider the cost function $c(x,y)=(x-y)^2/2$ instead of $(x-y)^2$. This is a nonnegative continuous function, so Theorem 2.3 applies.

A nice property of quadratic cost is being *invariant to translations*. If μ and ν are measures on $\mathbb{R}^d$ and z is an arbitrary vector in $\mathbb{R}^d$, we can translate ν by z by defining the probability measure $\nu_z(B)=\nu(B-z)$ for measurable B. A coupling $\pi_z\in\Pi(\mu,\nu_z)$ is easily obtained from a coupling $\pi\in\Pi(\mu,\nu)$ by setting $\pi_z(A\times B)=\pi(A\times(B-z))$. By the identity $c(x,y)=x^2/2+y^2/2-xy$ we have

$$\begin{aligned}I(\pi_v)&=\int_{\mathbb{R}^{2d}}c(x,y)d\pi_z(x,y)=\int_{\mathbb{R}^{2d}}c(x,y-z)d\pi(x,y)=\\&=\int_{\mathbb{R}^{2d}}c(x,y)d\pi(x,y)+\int_{\mathbb{R}^{2d}}(yz-xz+z^2/2)d\pi(x,y)=\\&=I(\pi)+\frac{z^2}{2}+z\left(\int_{\mathbb{R}^d}xd\mu(x)-\int_{\mathbb{R}^d}yd\nu(y)\right).\end{aligned}\tag{5.1}$$

Thus minimizing I over $\Pi(\mu,\nu_z)$ is equivalent to minimization over $\Pi(\mu,\nu)$. Denote the (vector) means of μ and ν

$$\int_{\mathbb{R}^d}xd\mu(x)=(a_1,\dots,a_d),\qquad\qquad\int_{\mathbb{R}^d}yd\nu(y)=(b_1,\dots,b_d),$$

then (5.1) is minimal when $z_i=a_i-b_i$. Therefore, if we are allowed to translate the measures before coupling them, the optimal translation is such that the mean vectors of

μ and ν are equal. In particular, if $\mu = \nu_z$ then the trivial map $x \mapsto x - z$ is the unique optimal transport map (and transference plan).
Let μ and ν be probability measures on $\mathbb{R}^d$ with finite second moments

$$M_2 = \int_{\mathbb{R}^d} \frac{x^2}{2} d\mu(x) + \int_{\mathbb{R}^d} \frac{y^2}{2} d\nu(y) < \infty.$$

The trivial inequality $(x-y)^2 \le 2x^2 + 2y^2$ shows that c is integrable for any $\pi \in \Pi(\mu, \nu)$. Recall that the Monge–Kantorovich problem is to find

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^{2d}} \frac{(x-y)^2}{2} d\pi(x,y) = M_2 - \sup_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^{2d}} xy d\pi(x,y).$$

The dual problem is to find integrable $\varphi \in L_1(\mu)$, $\psi \in L_1(\nu)$, maximizing

$$J(\varphi, \psi) = \int_{\mathbb{R}^d} \varphi d\mu + \int_{\mathbb{R}^d} \psi d\nu, \qquad (\varphi, \psi) \in \Phi_c,$$

where Φ_c is the set of pairs of functions such that for any x and any y,

$$\varphi(x) + \psi(y) \le \frac{(x-y)^2}{2} = \frac{x^2}{2} + \frac{y^2}{2} - xy.$$

Since $x \mapsto x^2/2$ and $y \mapsto y^2/2$ are in $L_1(\mu)$ and $L_1(\nu)$, we have that $(\varphi, \psi) \in \Phi_c$ if and only if $(\tilde\varphi, \tilde\psi) = (x^2/2 - \varphi, y^2/2 - \psi) \in \tilde\Phi_c$, where

$$\tilde\Phi_c = \left\{ (\tilde\varphi, \tilde\psi) : \tilde\varphi(x) + \tilde\psi(y) \ge xy \quad \forall x, y \right\}.$$

Since $J(\varphi, \psi) = M_2 - J(\tilde\varphi, \tilde\psi)$, maximization of J over Φ_c is equivalent to minimization over $\tilde\Phi_c$. Therefore, the Kantorovich duality

$$\sup_{(\varphi,\psi)\in\Phi_c} J(\varphi, \psi) = \inf_{\pi \in \Pi(\mu,\nu)} I(\pi), \qquad I(\pi) = \int_{\mathbb{R}^{2d}} \frac{(x-y)^2}{2} d\pi(x,y)$$

is equivalent to

$$\inf_{(\tilde\varphi,\tilde\psi)\in\tilde\Phi_c} J(\tilde\varphi, \tilde\psi) = \sup_{\pi \in \Pi(\mu,\nu)} \tilde I(\pi), \qquad \tilde I(\pi) = \int_{\mathbb{R}^{2d}} xy d\pi(x,y).$$

For $\varphi \in L_1(\mu)$ we define, with analogy to (2.4),

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

In the definition (2.4) measurability is not a-priori guaranteed. Fortunately, this is not the case here.

Proposition 5.1 *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be an arbitrary (perhaps not even measurable!) function. Then φ^* is convex and lower semi-continuous. In particular it is measurable.*

Proof. The convexity of φ^* follows immediately from the definitions, and its lower semi-continuity follows from the (lower semi-)continuity of $y \mapsto xy$. Indeed, suppose that $y_n \to y$ and let $x \in \mathbb{R}^d$, then $\varphi^*(y_n) \geq xy_n - \varphi(x)$, so

$$\liminf_{n\to\infty} \varphi^*(y_n) \geq \liminf_{n\to\infty} xy_n - \varphi(x) = xy - \varphi(x).$$

Taking the supremum over x, we find

$$\liminf_{n\to\infty} \varphi^*(y_n) \geq \sup_{x\in\mathbb{R}^d} xy - \varphi(x) = \varphi^*(y),$$

establishing lower semi-continuity.
It is easy to see that $\varphi^{***} = \varphi^*$. A pair $(\varphi^{**}, \varphi^*)$ (such that φ^* is not always infinite) is a pair of **proper convex conjugate functions**.
It has already been shown in Section 2 that an optimal transference plan π exists. Our goal now is to show that an optimal pair (φ, ψ) exists. Our strategy is to take a sequence (φ_k, ψ_k) that approximate the optimal value of J, and from this sequence construct an optimal pair. The problem is that $(\varphi_k - a_k, \psi_k + a_k)$ yields the same value for J, so we cannot expect the sequences (φ_k) and (ψ_k) to converge; the following lemma provides values of a_k such that these sequence are well-behaved. Recall that after to the algebraic manipulations above, J is to be minimized.

Lemma 5.2 *Let $X, Y \subseteq \mathbb{R}^d$, $\mu \in P(X)$ and $\nu \in P(Y)$ be probabilities with finite second moments*

$$M_2 \stackrel{\text{def}}{=} \int_X \frac{x^2}{2} d\mu(x) + \int_Y \frac{y^2}{2} d\nu(y) < \infty.$$

Let $\tilde{\Phi}$ be the set of integrable functions (φ, ψ) with values in $\mathbb{R} \cup \{\infty\}$, and such that $\varphi(x) + \psi(y) \geq xy$ for any $x \in X$, $y \in Y$. Consider a minimizing sequence $(\varphi_k, \psi_k) \in \tilde{\Phi}$ such that

$$\lim_{k\to\infty} J(\varphi_k, \psi_k) = \inf_{(\varphi,\psi)\in\tilde{\Phi}} J(\varphi, \psi) \stackrel{\text{def}}{=} J_0.$$

Then there exists a sequence of proper convex conjugate functions $(\overline{\varphi_k}, \overline{\psi_k}) = (\varphi_k^{**} - a_k, \varphi_k^* + a_k)$ such that
(1) for any x and any y,

$$\overline{\varphi_k}(x) \geq -\frac{x^2}{2} \qquad \text{and} \qquad \overline{\psi_k}(y) \geq -\frac{y^2}{2}.$$

(2) $\overline{\varphi_k}$ and $\overline{\psi_k}$ are integrable, and are also a minimizing sequence for J.

$$(3) \quad \limsup_{k\to\infty} \inf_{x\in X} \left(\overline{\varphi_k}(x) + \frac{x^2}{2} \right) \leq J_0 + M_2.$$

$$(4) \quad \limsup_{k\to\infty} \inf_{y\in Y} \left(\overline{\psi_k}(y) + \frac{y^2}{2} \right) \leq J_0 + M_2.$$

Proof. Since $\varphi_k^* \le \psi_k$ and $\varphi_k^{**} \le \varphi_k$, it follows that $J(\overline{\varphi_k}, \overline{\psi_k}) \le J(\varphi_k, \psi_k)$ if the left hand side is defined, and then $(\overline{\varphi_k}, \overline{\psi_k})$ is a minimizing sequence. Once we prove (1), the functions $\overline{\varphi_k} + x^2/2$ and $\overline{\psi_k} + y^2/2$ are nonnegative, so their integrals with respect to μ and ν are well defined, and

$$J\left(\overline{\varphi_k} + \frac{x^2}{2}, \overline{\psi_k} + \frac{y^2}{2}\right) \le M_2 + J(\varphi_k, \psi_k) \to M_2 + J_0, \qquad k \to \infty.$$

As $M_2 + J_0$ is finite, it follows that $\overline{\varphi_k}$ and $\overline{\psi_k}$ are integrable and are a minimizing sequence for J by the above.

To prove (1), observe that for some y_0, $\psi_k(y_0)$ is finite and therefore for any x, $\varphi_k(x) \ge xy_0 - b_0$, where $b_0 = \psi_k(y_0)$. This provides the upper bound

$$\varphi_k^*(y_0) = \sup_{x \in X} (xy_0 - \varphi_k(x)) \le b_0.$$

On the other hand, $\varphi_k(x_0)$ is finite for some x_0, so for any y, $\varphi_k^*(y) \ge x_0 y - d_0$, where $d_0 = \varphi_k(x_0)$. Now choose

$$a_k \stackrel{\text{def}}{=} \inf_{y \in Y} \left(\varphi_k^*(y) + \frac{y^2}{2}\right).$$

Since $\varphi_k^*(y_0) < \infty$, $a_k < \infty$. As φ_k^* is bounded from below by an affine function, $a_k > -\infty$, so it is finite. Set $\overline{\psi_k} = \varphi_k^* - a_k$ and $\overline{\varphi_k} = \varphi_k^{**} + a_k = (\overline{\psi_k})^*$. The remaining part of (1) now follows from $xy + x^2/2 \ge -y^2/2$:

$$\overline{\varphi_k}(x) + \frac{x^2}{2} = (\overline{\psi_k})^*(x) + \frac{x^2}{2} = \sup_{y \in Y} \left(xy - \overline{\psi_k}(y) + \frac{x^2}{2}\right) \ge$$

$$\ge \sup_{y \in Y} \left(-\frac{y^2}{2} - \overline{\psi_k}(y)\right) = a_k - \inf_{y \in Y} \left(\frac{y^2}{2} + \varphi_k^*(y)\right) = 0.$$

(3) and (4) follow from

$$J(\overline{\varphi_k}, \overline{\psi_k}) + M_2 = \int_X \left(\overline{\varphi_k}(x) + \frac{x^2}{2}\right) d\mu(x) + \int_Y \left(\overline{\psi_k}(y) + \frac{y^2}{2}\right) d\nu(y) \ge$$

$$\ge \inf_{x \in X} \left(\overline{\varphi_k}(x) + \frac{x^2}{2}\right) + \inf_{y \in Y} \left(\overline{\psi_k}(y) + \frac{y^2}{2}\right).$$

The result follows from the nonnegativity of the two infima and $J(\overline{\varphi_k}, \overline{\psi_k}) \to J_0$.

From the lemma it follows that when minimizing J, we need only consider $J(\varphi^{**}, \varphi^*)$ for $\varphi \in L_1(\mu)$. Since we can set φ to equal infinity outside the support of μ, there is no loss of generality in assuming $X = Y = \mathbb{R}^d$, but sometimes this selection is less convenient.

With the well-behavedness of the functions $\overline{\psi_k}$ and $\overline{\varphi_k}$, we can prove convergence of a minimizing sequence to a minimizer.

Proposition 5.3 (Optimal convex conjugate functions) *Let μ and ν be probabilities on $\mathbb{R}^d$ with finite second moments, then there exists a pair of lower semi-continuous convex conjugate functions $(\varphi^{**}, \varphi^*)$ such that $J(\varphi^{**}, \varphi^*) = J_0$.*

Proof. We prove the proposition only for the case where μ and ν are supported on compact sets $X, Y \subset \mathbb{R}^d$. The proof for the general case is in [14, p. 64–66].
Let (φ_k, ψ_k) be a minimizing sequence for J, and assume without loss of generality that they are of the form of Lemma 5.2, so that $\varphi_k^* = \psi_k$ and $\psi_k^* = \varphi_k$. By the definition of the $*$ operator, we see that $\{\psi_k\}$ are uniformly Lipschitz with the constant $\sup_{x\in X} \|x\|$ and $\{\varphi_k\}$ are uniformly Lipschitz with the constant $\sup_{y\in Y} \|y\|$. Furthermore, for large enough k there exist $x_k \in X$ and $y_k \in Y$ such that

$$-\sup_{x\in X} \frac{x^2}{2} \leq -\frac{x_k^2}{2} \leq \varphi_k(x_k) \leq J_0 + M_2 + 1$$

and

$$-\sup_{y\in Y} \frac{y^2}{2} \leq -\frac{y_k^2}{2} \leq \psi_k(y_k) \leq J_0 + M_2 + 1.$$

In view of the boundedness of X and Y and the uniform Lipschitz condition, it follows that φ_k and ψ_k are also uniformly bounded. The uniform Lipschitz continuity implies equicontinuity, and the Arzelà–Ascoli theorem can be invoked. Up to extracting a subsequence, both (φ_k) and (ψ_k) converge uniformly to continuous functions on X and Y, φ and ψ. By the uniform convergence

$$J_0 = \lim_{k\to\infty} J(\varphi_k, \psi_k) = J(\varphi, \psi),$$

so it is a minimizing pair. Extend φ and ψ to equal infinity outside X and Y, and set

$$\varphi^*(y) = \sup_{x\in\mathbb{R}^d}(xy - \varphi(x)) \qquad \text{and} \qquad \varphi^{**}(x) = \sup_{y\in\mathbb{R}^d}(xy - \varphi^*(y)).$$

Then the pair $(\varphi^{**}, \varphi^*)$ is a minimizing pair of proper convex conjugate functions with $\varphi^{**}(x) + \varphi^*(y) \geq xy$ for any x and any y.

Now that the existence of solutions for both the primal and the dual problem is established, we turn to unveil connections between the solutions of the problem. These connections will also allow us to prove uniqueness of the solution of the dual, and uniqueness of the minimizing pair up to constant. Convexity, as we shall see, will play a key role here.
For a convex function $\varphi : X \to \mathbb{R} \cup \{\infty\}$ we define its domain by the set of point at which φ is finite. The domain is a convex set and as such, its boundary has Lebesgue measure zero (in fact, if $X \subseteq \mathbb{R}^d$ then $\partial dom(\varphi)$ is a set of Hausdorff dimension of at most $d-1$).
For a convex function φ, its **subdifferential** at $x \in X$ is the set

$$\partial\varphi(x) \stackrel{\text{def}}{=} \{y \in \mathbb{R}^d : \forall z \in X \quad \varphi(z) \geq \varphi(x) + y(z-x)\}.$$

For $x \in int(dom(\varphi))$, $\partial\varphi(x)$ is nonempty (by the Hahn–Banach theorem). φ is differentiable at x if and only if $\nabla\varphi = \{\partial\varphi(x)\}$, which geometrically means that the tangent plane lies below the graph of φ. In $int(dom(\varphi))$, φ is differentiable almost anywhere with

respect to Lebesgue measure (in fact, the set of points where φ is not differentiable is of Hausdorff dimension of at most $d-1$). A proof of all these statements can be found in [12].

We will identify the subdifferential of φ with its graph $\{(x,y) : y \in \partial\varphi(x)\}$. This graph is always a closed set, because if $x_n \to x$, $y_n \to y$ and $y_n \in \partial\varphi(x_n)$, then it immediately follows from the definitions that $y \in \partial\varphi(x)$. The relation between the subdifferential and the function φ^* is this: by definition, for any x and any y we have $\varphi(x) + \varphi^*(y) \geq xy$. The crucial point is when equality holds.

$$\varphi(x) + \varphi^*(y) = xy \iff \varphi(x) + \varphi^*(y) \leq xy \iff \forall z \quad \varphi(x) + yz - \varphi(z) \leq xy \iff$$

$$\iff \forall z \quad \varphi(z) \geq \varphi(x) + y(z-x) \iff y \in \partial\varphi(x).$$

With these tools at hand, we can prove the beautiful result for quadratic cost. Recall that the support of a measure $\sigma \in P(X)$ is the smallest closed set A such that $\sigma(A) = 1$. It is denoted by $supp(\sigma)$.

Theorem 5.4 (Optimal transportation for quadratic cost) *Let μ and ν be probability measures on $\mathbb{R}^d$ with finite second moments, and consider the cost function $c(x,y) = (x-y)^2$. Then*
(1) $\pi \in \Pi(\mu,\nu)$ is optimal if and only if it is supported on the graph of a subdifferential of a convex lower semi-continuous function φ.

$$supp(\pi) \subseteq Graph(\partial\varphi).$$

Equivalently,

$$\pi\left(\{(x,y) : y \in \partial\varphi(x)\}\right) = 1.$$

In addition, the pair (φ, φ^*) is a minimizer of J over $\tilde{\Phi}$, and $(x^2/2 - \varphi, y^2/2 - \varphi^*)$ is a maximizer of J over Φ.
(2) If μ is absolutely continuous with respect to Lebesgue measure, then there exists a unique (up to μ-measure zero set) gradient of a convex function, $\nabla\varphi$, that pushes μ forward to ν: $\nabla\varphi\#\mu = \nu$. $\nabla\varphi$ is the unique solution to the Monge problem, and the transference plan it induces,

$$\pi = (id \times \nabla\varphi)\#\mu \qquad \pi(A \times B) = \mu\left(A \cap (\nabla\varphi)^{-1}(B)\right),$$

is the unique solution to the Kantorovich problem.

Remark 9 *In fact it suffices that $\mu(A) = 0$ for any set A with Hausdorff dimension $d-1$.*

Proof. Firstly, $Graph(\partial\varphi)$ is a closed set, so the two statements of (1) are equivalent. The assumptions of the theorem imply that optimal solutions exist for both problems. Let

π be optimal for the primal and (φ, φ^*) be optimal for the dual, then by the Kantorovich duality and Proposition 1.1

$$\int_{\mathbb{R}^{2d}} xyd\pi(x,y) = J(\varphi,\varphi^*) = \int_{\mathbb{R}^{2d}} (\varphi(x)+\varphi^*(y))d\pi(x,y).$$

It follows that the integral of the nonnegative function $\varphi(x)+\varphi^*(y)-xy$ vanishes, and $\varphi(x)+\varphi^*(y) = xy$ π-almost anywhere. Equivalently, $y \in \partial\varphi(x)$ π-almost anywhere. Reversing the arguments proves the converse.

To prove (2), let φ and ψ be two lower semi-continuous convex functions such that $\nabla\varphi\#\mu = \nu = \nabla\psi\#\mu$. Since μ is absolutely continuous, both φ and ψ are differentiable μ-almost anywhere. By (1) the measures

$$\pi(A\times B) = \mu(A\cap(\nabla\varphi)^{-1}(B)) \qquad \text{and} \qquad \rho(A\times B) = \mu(A\cap(\nabla\psi)^{-1}(B))$$

are both optimal, since π is supported on $Graph(\partial\varphi) = \{(x, \nabla\varphi(x))\}$ and ρ on $Graph(\partial\psi)$. In particular, both pairs (φ,φ^*) and (ψ,ψ^*) are optimal and yield the same value for J. To see that $\nabla\varphi = \nabla\psi$, observe that

$$\int_{\mathbb{R}^{2d}} (\psi(x)+\psi^*(\nabla\varphi(x)))d\mu(x) = \int_{\mathbb{R}^{2d}} (\psi(x)+\psi^*(y))d\pi(x,y) = J(\psi,\psi^*) = J(\varphi,\varphi^*) =$$

$$= \int_{\mathbb{R}^{2d}} (\varphi(x)+\varphi^*(y))d\pi(x,y) = \int_{\mathbb{R}^{2d}} xyd\pi(x,y) = \int_{\mathbb{R}^d} x\nabla\varphi(x)d\mu(x).$$

The integral of the positive function

$$\psi(x)+\psi^*(\nabla\varphi(x)) - x\nabla\varphi(x)$$

equals zero, so this function vanishes μ-almost anywhere. This means that μ-almost anywhere, $\nabla\varphi(x) \in \partial\psi(x) = \{\nabla\psi(x)\}$, so $\nabla\varphi = \nabla\psi$ μ-almost anywhere. Under the conditions of μ any optimal transference plan π is given by $\pi = (id\times\nabla\varphi)\#\mu$, which proves the uniqueness of π. Since $\nabla\varphi$ is unique, φ is unique up to an additive constant (and a μ-measure zero set).

5.2 Other cost functions

The case of quadratic cost is special because of the decomposition $(x-y)^2 = x^2+y^2-2xy$, which allows us to use tools from convex analysis. In order to generalize Theorem 5.4 to other costs, we need to get back to the original dual problem: maximize J over

$$\Phi = \{(\varphi,\psi) : \varphi(x)+\psi(y) \le (x-y)^2\}.$$

We solved the equivalent problem, which is to minimize J over

$$\tilde{\Phi} = \{(\varphi,\psi) : \varphi(x)+\psi(y) \ge xy\}.$$

The relation between the solutions is that (φ, ψ) is optimal for $\tilde{\Phi}$ if and only if $(x^2/2 - \varphi, y^2/2 - \psi)$ is optimal for Φ. If μ is absolutely continuous, then there exists a unique (up to additive constant) φ such that (φ, φ^*) is optimal for $\tilde{\Phi}$ and, equivalently, the pair $(\psi, \psi^c) = (x^2/2 - \varphi, y^2/2 - \varphi^*)$ is optimal for Φ (the reader should verify that $(x^2/2 - \varphi)^c = y^2/2 - \varphi^*$). Furthermore, the unique solution to the Kantorovich problem is the transport map $\nabla\varphi$, which equals $x - \nabla\psi$.

Gangbo and McCann [6] generalize this result for convex cost functions of the form $c(x, y) = h(|x - y|)$ where h is strictly convex. Under some conditions on h (e.g. superlinear growth) they show that when μ is absolutely continuous, a unique solution to the Monge–Kantorovich problem is the map $x \mapsto x - (\nabla h)^{-1}(\nabla\varphi(x))$ (∇h is invertible because h is strictly convex). When $h(x) = x^2/2$, ∇h is the identity map and Theorem 5.4 is recovered.

When h is concave, c is a metric, and therefore all shared mass can stay in place. If h is strictly concave, then c satisfies the strict triangle inequality (Lemma 2.8) and Gangbo and McCann show that all shared mass *has* to stay in place. Therefore, the problem can be reduced to the case where μ and ν have disjoint supports. In this case, if μ is absolutely continuous then, again, the unique optimal solution is of the form $x \mapsto x - (\nabla h)^{-1}(\nabla\varphi(x))$. It is using this theorem that the functions in Example 1 were found.

If μ and ν have shared mass then it is possible that no optimal transport map exists: let μ be Lebesgue measure on $[0, 1]$ and ν be half Lebesgue measure on $[0, 2]$. After leaving all shared mass in place, the problem is to map $[0, 1]$ to $[1, 2]$ (which is carried out optimally by $s(x) = 2 - x$). Therefore, any point in $[0, 1]$ has to be split; half of it stays in place and the other half is moved by s. Uniqueness implies that there is no optimal transport map.

Unlike the proof presented here for quadratic cost, Gangbo and McCann impose no moment restrictions on the measures μ and ν; they only assume that some transference plan with finite cost exists. The tools used in their proof are c-concavity, c-superdifferential and cyclical monotonicity, the first two being generalizations of convexity and the subdifferential of a convex function. These notions will be surveyed in Section 7.

6 A gradient descent algorithm for quadratic cost

With existence and uniqueness established for a rather large family of cost functions, in this section we address a more practical question: how to find the optimal transference plan, which is in fact a transport map? When the cost function is quadratic $c(x,y) = (x-y)^2$, the relationship established in Section 5 between the Monge–Kantorovich problem and convexity gives rise to a simple gradient descent algorithm, which is presented in this section. It is due to Chartrand et al. [4].

In this setting, X and Y are compact subsets of $\mathbb{R}^d$, and μ and ν are absolutely continuous probabilities on X and Y with densities f_1 and f_2. For a function $\varphi : X \to \mathbb{R} \cup \{\infty\}$ we defined its convex conjugate $\varphi^* : Y \to \mathbb{R}$ by

$$\varphi^*(y) = \sup_{x \in X}(xy - \varphi(x)), \qquad xy = \langle x, y\rangle = \sum_{i=1}^{d} x_i y_i.$$

Under the above conditions Theorem 5.4 applies: the functional L defined by

$$L(\varphi,\psi) = \int_X \varphi d\mu + \int_Y \psi d\nu = \int_X f_1(x)\varphi(x)dx + \int_Y f_2(y)\psi(y)dy$$

defined on the set of functions

$$\tilde{\Phi} \stackrel{\text{def}}{=} \{(\varphi,\psi) : \varphi(x) + \psi(y) \geq xy \ \forall x, y\}$$

admits a minimizing pair (φ,ψ) of convex conjugate lower semi-continuous functions: $\psi = \varphi^*$ and $\varphi = \psi^*$. Furthermore, the pair is unique up to an additive constant (and a μ-measure zero set). The gradient $\nabla\varphi$ is the optimal transport map from μ to ν.

It follows from the above that the constrained maximization problem can be recast as an unconstrained problem. Define

$$M(\varphi) = L(\varphi,\varphi^*) = \int_X f_1(x)\varphi(x)dx + \int_Y f_2(y)\varphi^*(y)dy, \qquad \varphi \in C(X),$$

where $C(X)$ is the set of continuous real-valued functions on X. Since $(\varphi,\varphi^*) \in \tilde{\Phi}$, we are looking for a maximizer of M with no constraints. The main result in [4] is

Theorem 6.1 *Let X and Y be compact sets in $\mathbb{R}^d$, $f_1 \in L_1(\mu)$ and $f_2 \in L_1(\nu)$ be non-negative integrable functions. Consider the functional $M : C(K_1) \to \mathbb{R}$ defined by*

$$M(\varphi) = \int_X f_1(x)\varphi(x)dx + \int_Y f_2(y)\varphi^*(y)dy.$$

Then M is convex, Lipschitz and Hadamard differentiable, with Hadamard derivative

$$M'(\varphi) = f_1 - (f_2 \circ \nabla\varphi^{**})det(D^2\varphi^{**}).$$

Furthermore, M has a unique (up to an additive constant) convex minimizer φ. If f_1 and f_2 are densities ($\|f_1\|_1 = 1 = \|f_2\|_1$), then $s = \nabla\varphi$ is the unique solution to the Monge–Kantorovich problem.

Remark 10 *If $\varphi = \psi^*$ for some ψ, then $\varphi^{**} = \varphi$, and the Hadamard derivative is*

$$M'(\varphi) = f_1 - (f_2 \circ \nabla\varphi) det(D^2\varphi).$$

A gradient descent algorithm can then be applied with the iteration step

$$\varphi_{n+1} = \varphi_n - \alpha_n M'(\varphi_n).$$

When $\varphi = \varphi^{**}$ and the derivative of M vanishes, we have

$$f_1 = (f_2 \circ \nabla\varphi) det(D^2\varphi).$$

Then a change of variables shows that $\nabla\varphi$ pushes μ (with density f_1) forward to ν (with density f_2).

Proof of Theorem 6.1 (sketch) We only have to prove that M is convex, Lipschitz and differentiable, since the other statements have been proven earlier. We wish to show that

$$M(t\varphi + (1-t)\psi) \leq tM(\varphi) + (1-t)M(\psi), \qquad \varphi, \psi \in C(X), \quad t \in [0,1].$$

Evaluating both sides, the inequality boils down to

$$\int_Y \left(\sup_{x \in X} (xy - t\varphi(x) - (1-t)\psi(x)) \right) f_2(y)dy \leq$$

$$\leq \int_Y \left(\sup_{x,z \in X} (txy - t\varphi(x) + (1-t)zy - (1-t)\psi(z)) \right) f_2(y)dy,$$

which follows from $f_2 \geq 0$ and the specification $x = z$.

To show that M is Lipschitz, observe firstly that any continuous function on X is bounded. For $\varphi, \psi \in C(X)$ we have

$$\|\varphi^* - \psi^*\|_\infty \leq \|\varphi - \psi\|_\infty.$$

Thus

$$|M(\varphi) - M(\psi)| \leq \int_X |\varphi - \psi| f_1 + \int_Y |\varphi^* - \psi^*| f_2 \leq \|\varphi - \psi\|_\infty \left(\int_X f_1 + \int_Y f_2 \right),$$

and M is Lipschitz with the constant $(\|f_1\|_1 + \|f_2\|_1)$.

The more difficult part is the differentiability of M. We begin with a directional derivative: let $u, v \in C(X)$, then

$$D_v M(u) = \lim_{t \to 0^+} \frac{M(u+tv) - M(u)}{t} = \int_X v f_1 + \lim_{t \to 0^+} \frac{1}{t} \left(\int_Y ((u+tv)^* - u^*) f_2 \right).$$

Since u^* is a convex function, it is differentiable almost anywhere. The reader can verify that it is valid to interchange the limit and the integral.

Fix $y \in Y$ such that u^* is differentiable at y, and set $x_0 = \nabla u^*(y)$. Then $u^*(y) \geq xy - u(x)$, with equality if and only if $x = x_0$. For any t choose $x_t \in (\partial(u+tv)^*)(y)$ (which is

nonempty as $(u+tv)^*$ is convex), then $(u+tv)^*(y) \geq xy - u(x) - tv(x)$, with equality when $x = x_t$. Combining the bounds

$$\frac{(u+tv)^*(y) - u^*(y)}{t} \leq \frac{x_t y - u(x_t) - tv(x_t) + u(x_t) - x_t y}{t} = -v(x_t),$$

and

$$\frac{(u+tv)^*(y) - u^*(y)}{t} \geq \frac{x_0 y - u(x_0) - tv(x_0) - u(x_0) + x_0 y}{t} = -v(x_0),$$

we obtain

$$0 \leq \frac{(u+tv)^*(y) - u^*(y)}{t} + v(x_0) \leq v(x_0) - v(x_t). \tag{6.1}$$

As v is bounded, $(u+tv)^*(y) \to u^*(y)$ and in addition $tv(x_t) \to 0$ when $t \to 0$. Therefore, if $(x_{t_n}) \to x$ for a subsequence $t_n \to 0$, then

$$u^*(y) = \lim_{n\to\infty} (u + t_n v)^*(y) = \lim_{n\to\infty} x_{t_n} y - u(x_{t_n}) - t_n v(x_{t_n}) = xy - u(x).$$

Since the only element of X satisfying the above equality is x_0, we conclude that the only partial limit of the sequence (x_t) can be x_0, and as X is compact, $x_t \to x_0$. Combining this convergence with (6.1), we have

$$\lim_{t\to 0^+} \frac{(u+tv)^*(y) - u^*(y)}{t} = -v(x_0) = -v(\nabla u^*(y)). \tag{6.2}$$

Equation (6.2) holds for any y where u^* is differentiable, so it holds almost anywhere. Thus

$$D_v M(u) = \int_X v f_1 - \int_Y (v \circ \nabla u^*) f_2.$$

Changing of variables $y = \nabla\varphi^{**}(x)$ in the second integral (see [4] and the references therein for justification) gives

$$D_v M(u) = \int_X v f_1 - \int_{(\varphi^{**})^{-1}(Y)} (v \circ \nabla u^* \circ \nabla u^{**})(f_2 \circ \nabla\varphi^{**}) det(D^2\varphi^{**}).$$

Next, observe that $u^{***} = u^*$, both u^* and u^{**} are differentiable almost anywhere, and

$$y = \nabla u^{**}(x) \iff u^{***}(y) + u^{**}(x) = xy \iff u^*(y) + u^{**}(x) = xy \iff x = \nabla\varphi(y).$$

In other words, $\nabla u^* \circ \nabla u^{**} = id$ and as $(\varphi^{**})^{-1}(Y) = X$ (see [4]), the directional derivative simplifies to

$$D_v M(u) = \int_X (f_2 \circ \nabla\varphi^{**}) det(D^2\varphi^{**}) v.$$

As explained in [4], this suffices for the existence of the Hadamard derivative.

7 Characterization of optimal couplings by their support

In this section, we introduce three important notions: c-monotonicity, c-concavity and c-superdifferential. When c is quadratic cost, these notions are closely related to convexity in the usual sense. We establish another characterization of the optimal transport plans, as c-monotone sets. As an important consequence, we prove an elegant stability result for the Kantorovich problem, Theorem 7.12, due to Schachermayer and Teichmann [13].

7.1 c-monotonicity

Consider the discrete Monge–Kantorovich problem, where

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, \qquad \nu = \frac{1}{n}\sum_{j=1}^{n}\delta_{y_j}, \qquad c(x_i, y_j) = c_{ij},$$

and assume that the indices are arranged so that the identity permutation is optimal. Then for any permutation $\sigma \in S_n$ we have

$$\sum_{i=1}^{n} c_{ii} \le \sum_{i=1}^{n} c_{i,\sigma(i)}. \tag{7.1}$$

In fact, the identity permutation is optimal if and only if (7.1) holds for any permutation $\sigma \in S_n$.

By writing σ as disjoint cycles, it is easy to see that (7.1) is equivalent to the assertion that for any $m \le n$ and distinct indices $i_1, \dots, i_m$,

$$\sum_{k=1}^{m} c_{i_k,i_k} \le \sum_{k=1}^{m} c_{i_k,i_{k-1}}, \qquad i_0 \stackrel{\text{def}}{=} i_m. \tag{7.2}$$

We now wish to show that a variant of (7.2) holds for arbitrary measures μ and ν.

Definition 7.1 *A set $S \subseteq X \times Y$ is **c-monotone** if for any collection of finitely many points $(x_i, y_i) \in S$, $i = 1, \dots, n$,*

$$\sum_{i=1}^{n} c(x_i, y_i) \le \sum_{i=1}^{n} c(x_i, y_{i-1}), \qquad y_0 = y_n. \tag{7.3}$$

A measure $\pi \in P(X \times Y)$ is c-monotone if there exists a c-monotone S such that $\pi(S) = 1$.

If μ and ν are discrete measures and σ is an optimal permutation for the Kantorovich problem, then the coupling $\pi = (1/n)\sum_{i=1}^{n} \delta\{(x_i, y_{\sigma(i)})\}$ is c-monotone. In fact, even if the optimal permutation is not unique, the set

$$S = \{(x_i, y_{\sigma(i)}) : i = 1, \dots, n, \sigma \in S_n \text{ optimal}\}$$

is c-monotone. Furthermore, $\pi(S) = 1$ for any optimal π, and π is optimal if and only if it is c-monotone. The following proposition shows the "only if" for arbitrary measures, when c is continuous. The more difficult "if" will be proved in the next subsection. Recall that the support of a probability measure π ($supp(\pi)$) is the smallest closed set A such that $\pi(A) = 1$.

Proposition 7.2 *Let X and Y be Polish spaces and suppose that the cost function c is nonnegative and continuous. For $\mu \in P(X)$ and $\nu \in P(Y)$, suppose that π is an optimal coupling between μ and ν, then $supp(\pi)$ is c-monotone. In particular, π is c-monotone.*

Proof. We present the proof of Gangbo and McCann [6]. Suppose that the converse is true, then there exist n points (x_i, y_i) in the support of π such that

$$f(x_1, \ldots, x_n, y_1, \ldots, y_n) \stackrel{\text{def}}{=} \sum_{i=1}^{n} c(x_i, y_{i-1}) - c(x_i, y_i) < -2\varepsilon < 0, \qquad (y_0 = y_n).$$

Since $(x_i, y_i) \in supp(\pi)$, in some sense π "transports x_i to y_i". The idea is to modify π by constructing a new measure π' that "transports x_i to y_{i-1}" and consequently incur a smaller cost than the cost of π. As f is continuous, there exist compact neighbourhoods $U_i \subset X$ and $V_i \subset Y$ with $(x_i, y_i) \in U_i \times V_i$ and $f < -\varepsilon$ on $(\prod_{i=1}^n U_i) \times (\prod_{i=1}^n V_i)$. For brevity, we denote $Z_i = U_i \times V_i$. Since $(x_i, y_i) \in supp(\pi)$, $\lambda \stackrel{\text{def}}{=} \min_i \pi(Z_i)$ is strictly positive. In addition, we can assume without loss of generality that Z_i are mutually disjoint. For any i define $\pi_i \in P(X \times Y)$ to be the probability measure supported on Z_i:

$$\pi_i(A) = \frac{\pi(Z_i \cap A)}{\pi(Z_i)}, \qquad A \subseteq X \times Y \text{ measurable.}$$

(π_i is called *the normalized restriction* of π to Z_i) As the Z_i are disjoint, $\pi - \lambda \sum_{i=1}^n \pi_i$ is still a positive measure. In order to "transport x_i to y_{i-1}", we need to couple the probability measures $(\pi_i)_{i=1}^n$ together. To this end, define the Borel probability space (Ω, τ) to be the product space of (Z_i, π_i), and let $\rho_i : \Omega \to Z_i$ be the projection map. As Z_i is itself a product space, we write $\rho_i = (\mathbf{u_i} \times \mathbf{v_i})$, where $\mathbf{u_i} : \Omega \to U_i$ and $\mathbf{v_i} : \Omega \to V_i$. Then $\pi_i = \rho_i \# \tau$; τ has marginals $\pi_1, \ldots, \pi_n$. The positive measure on $X \times Y$

$$\alpha_i = (\mathbf{u_i} \times \mathbf{v_{i-1}}) \# \tau, \qquad \mathbf{v_0} = \mathbf{v_n}$$

satisfies

$$\alpha_i(A \times Y) = \tau({\mathbf{u_i}}^{-1}(A)) = \pi_i(A \times Y), \qquad A \subseteq X, \tag{7.4}$$

and ($\pi_0 = \pi_n$)

$$\alpha_i(X \times B) = \tau({\mathbf{v_{i-1}}}^{-1}(B)) = \pi_{i-1}(X \times B), \qquad B \subseteq Y. \tag{7.5}$$

In addition (x_i, y_{i-1}) is in the support of α_i. α_i couples the X marginal of π_i with the Y marginal of π_{i-1}. Finally, we let

$$\pi' = \pi + \lambda \sum_{i=1}^{n} (\alpha_i - \pi_i).$$

By (7.4) and (7.5), π' has the same marginals as π. It is positive because $\pi - \lambda \sum_{i=1}^{n} \pi_i$ is a positive measure. But by the hypothesis on f and a change of variables,

$$C(\pi') - C(\pi) = \lambda \int_{\Omega} f(\mathbf{u_1}, \ldots, \mathbf{u_n}, \mathbf{v_1}, \ldots, \mathbf{v_n}) d\tau < -\lambda\varepsilon < 0.$$

Thus π is not optimal.

It is not difficult to strengthen Proposition 7.2 and prove existence of a c-monotone set S that includes the support of *any* optimal transference plan π: take $S = \cup supp(\pi)$ for π optimal.

When X and Y are Euclidean spaces and the cost function $c(x, y) = \|x - y\|^2 = (x - y)^2$ is quadratic cost, a c-monotone set is called **cyclically monotone**. It turns out that cyclically monotone sets are closely related to convexity: easy algebra shows that for quadratic cost, (7.3) is equivalent to

$$\sum_{k=1}^{m} y_{i_k}(x_{i_{k+1}} - x_{i_k}) \leq 0, \tag{7.6}$$

where $i_{m+1} \stackrel{\text{def}}{=} i_1$. Similarly to the identity $(x - y)^2 = x^2 + y^2 - 2xy$ used in Section 5, this equivalence and the following theorem provide the connection to convexity, by characterizing cyclically monotone sets as gradients of convex functions. Together with Proposition 7.2, it can be used to prove the theorem for quadratic cost (Theorem 5.4), without even assuming finite second moments.

Theorem 7.3 (Rockafellar) *A nonempty $\Gamma \subseteq \mathbb{R}^{2d}$ is cyclically monotone if and only if it is included in the graph of the subdifferential of a proper lower semi–continuous convex function.*

Proof. By definition, the subdifferential of a convex function φ at x is

$$\partial\varphi(x) = \{y \in \mathbb{R}^d : \forall z \ \varphi(z) \geq \varphi(x) + y(z - x)\}.$$

If φ is convex and $y_i \in \partial\varphi(x_i)$ for $i = 1, \ldots, n$ then specifying $z = x_{n+1}$ gives

$$\forall i \quad \varphi(x_{i+1}) \geq \varphi(x_i) + y_i(x_{i+1} - x_i), \qquad x_{n+1} = x_1.$$

Summing up these n inequalities, one recovers (7.6). Thus $Graph(\partial\varphi)$ is cyclically monotone, and so is any of its subsets.

For the converse, suppose that Γ is cyclically monotone and nonempty. Let $(x_0, y_0) \in \Gamma$ and define

$$\varphi(x) = \sup_{m \in \mathbb{N}} \Big(\sup\{y_m(x - x_m) + y_{m-1}(x_m - x_{m-1}) + \cdots + y_0(x_1 - x_0) : (x_m, y_m) \in \Gamma\} \Big). \tag{7.7}$$

Then $\varphi(x_0) \leq 0$ because Γ is cyclically monotone, so Γ is not infinite everywhere. This is the only part where we use Γ being cyclically monotone. Indeed, it is easy to see that φ

is identically infinite if Γ is not cyclically monotone, by taking (x_i, y_i) that violate (7.6) and repeatedly plugging them in the definition of φ. As a supremum of affine functions, φ is convex and lower semi-continuous, and we only need to show that Γ is included in the subdifferential of φ. Given $(x^*, y^*) \in \Gamma$, we need to show that for an arbitrary z,

$$\varphi(z) \geq y^*(z - x^*) + \varphi(x^*).$$

It is sufficient to show that whenever $(x_i, y_i) \in \Gamma$ for $i = 1, \ldots, m$,

$$\varphi(z) \geq y^*(z - x^*) + y_m(x^* - x_m) + y_{m-1}(x_m - x_{m-1}) + \cdots + y_0(x_1 - x_0),$$

which is obvious from the definition of φ.

7.2 Strong c-monotonicity

In this subsection we introduce a strengthened version of c-monotonicity which we call, following [13], strong c-monotonicity. We prove that π is optimal if and only if it is *strongly* c-monotone, suggesting that the latter is the "correct" notion in our context. In addition, we show that a c-monotone transference plan is strongly c-monotone. Some restrictions on the cost function have to be imposed, but they are rather weak: all the statements here hold true if c never takes infinite values.

Definition 7.4 *A Borel set $\Gamma \subseteq X \times Y$ is **strongly c-monotone** if there exist Borel functions $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ and $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ such that for any x and any y, $\varphi(x) + \psi(y) \leq c(x, y)$ with equality if $(x, y) \in \Gamma$.*
$\pi \in P(X \times Y)$ is strongly c-monotone if there exists a strongly c-monotone Γ with $\pi(\Gamma) = 1$.

It is clear that a strongly c-monotone set is c-monotone: if $(x_i, y_i) \in \Gamma$ then

$$\sum_{i=1}^{n} c(x_i, y_{i-1}) \geq \sum_{i=1}^{n} (\varphi(x_i) + \psi(y_{i-1})) = \sum_{i=1}^{n} (\varphi(x_i) + \psi(y_i)) = \sum_{i=1}^{n} c(x_i, y_i).$$

If π is strongly c-monotone and the functions φ and ψ are integrable, then

$$I(\pi) \stackrel{\text{def}}{=} \int_{X \times Y} c d\pi = \int_X \varphi d\mu + \int_Y \psi d\nu \stackrel{\text{def}}{=} J(\varphi, \psi), \qquad (\varphi, \psi) \in \Phi_c,$$

so π is optimal for the primal and the pair (φ, ψ) for the dual. It should therefore not be surprising that strongly c-monotone transference plans are optimal. The proof, however, is subtle.
We also need to extend the notions of convex conjugates and subdifferential as defined in Section 5:

$$\varphi^*(x) = \sup_{y \in Y} (xy - \varphi(y)),$$

and

$$\partial\varphi(x) = \{y \in Y : \forall z \in X \quad \varphi(z) \geq \varphi(x) + y(z - x)\}.$$

Definition 7.5 *A function $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ is **c-concave** if there exists a function $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ such that*

$$\varphi(x) = \psi^c(x) \stackrel{\text{def}}{=} \inf_{y \in Y} (c(x,y) - \psi(y)).$$

A c-concave function on Y is defined analogously.
*Given $c : X \times Y \to \mathbb{R}_+ \cup \{\infty\}$ and $\varphi : X \to \mathbb{R} \cup \{-\infty\}$, the **$c$-superdifferential** of φ at x is*

$$\partial_c \varphi(x) \stackrel{\text{def}}{=} \{y \in Y : \forall z \in X \quad \varphi(z) \leq \varphi(x) + c(z,y) - c(x,y)\}.$$

The c-superdifferential of a function on Y is defined analogously as a subset of X. We will identify the c-superdifferential of a function with its graph $\{(x,y) : y \in \partial_c \varphi\}$.

When $c(x,y) = (x-y)^2/2$, $\varphi = \psi^c$ if and only if $(x^2/2 - \varphi) = (y^2/2 - \psi)^*$, and in addition $y \in \partial_c \varphi(x)$ if and only if $y \in \partial \tilde{\varphi}(x)$, $\tilde{\varphi}(x) = x^2/2 - \varphi(x)$. The analogue of the equivalence $\varphi(x) + \varphi^*(y) = xy$ if and only if $y \in \partial\varphi$ also holds in this more general context: since $\psi^c(x) + \psi(y) \leq c(x,y)$ holds for any x and any y, we have

$$\psi^c(x) + \psi(y) = c(x,y) \iff x \in \partial_c \psi(y).$$

(This will be proved in the next proposition.)
The strategy in this subsection is as follows: firstly, we show that c-monotone transference plans are strongly c-monotone, when c is finitely valued. Then we prove that optimal transference plans are strongly c-monotone, a stronger result than Proposition 7.2. Next, we show the converse—that any strongly c-monotone π is optimal, and the stability theorem follows as a corollary, due to the stability of c-monotonicity under weak convergence: when $\pi_n \to \pi$ and π_n is c-monotone, so is π.

Proposition 7.6 *Let X and Y be Polish spaces, $\pi \in P(X \times Y)$ and $c : X \times Y \to \mathbb{R}_+$ a lower semi-continuous function such that π is c-monotone. Then there exist Borel functions $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ and $\psi : X \to \mathbb{R} \cup \{-\infty\}$ such that*
(1) $\pi(\partial_c \varphi) = 1$.
(2) $\psi = \varphi^c$ ν-almost surely (here ν is the Y-marginal of π).
(3) For any x, y, $\varphi(x) + \psi(y) \leq c(x,y)$, with equality holding π-almost surely.
In particular, π is strongly c-monotone.

Proof. Let Γ be a Borel c-monotone set with $\pi(\Gamma) = 1$. By Lusin's theorem, we may assume without loss of generality that $\Gamma = \bigcup \Gamma_k$ where Γ_k are compact and c is continuous on Γ_k. In order to build φ, we use the analogue of (7.7): fix some $(x_0, y_0) \in \Gamma$, and let

$$\varphi(x) = \inf_{p \in \mathbb{N}} \left(\inf \left\{ c(x,y_p) - c(x_p,y_p) + c(x_p,y_{p-1}) - c(x_{p-1},y_{p-1}) + \cdots + c(x_1,y_0) - c(x_0,y_0) : \right.\right.$$

$$(x_i, y_i) \in \Gamma, i = 1, \ldots, p\Big\}\Big).$$

As in the proof of Theorem 7.3, It is straightforward to check that $\Gamma \subseteq \partial_c\varphi$, and $\varphi(x_0) = 0$ so φ is not identically negative infinite. Set

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

Ambrosio and Pratelli [1, Theorem 3.2, Steps 1 and 2] show that φ and φ^c are measurable, and φ^c is real valued up to a ν-measure zero set (they also prove optimality of the pair (φ, φ^c) for the dual problem, under an additional weak condition on c). We may change φ^c on a ν-measure zero set (to equal minus infinity, say) to obtain a function $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ such that for all x and y,

$$\varphi(x) + \psi(y) \leq c(x, y). \tag{7.8}$$

Thus (1) and (2) are proved. It only remains to verify that equality holds in (7.8) when $(x, y) \in \Gamma$. But when $(x, y) \in \Gamma$, by (1) $(x, y) \in \partial_c\varphi$, so that for all $z \in X$,

$$\forall z \in X, \qquad \varphi(z) \leq \varphi(x) + c(z, y) - c(x, y).$$

Thus

$$c(x, y) - \varphi(x) \leq \inf_{z\in X}(c(z, y) - \varphi(z)) = \varphi^c(y),$$

and the reverse inequality holds for any x, y. Therefore equality in (7.8) holds π-almost surely, completing the proof.
When c is allowed to take infinite values, there exist c-monotone sets that are not strongly c-monotone, see [13, Example 1].

The next step is to show that optimality implies strong monotonicity. To proceed, we need two lemmas that we state without proof. The first lemma asserts that if $f_n + g_n$ converges, then there exist real numbers r_n such that up to a extracting a subsequence, $\{f_n + r_n\}$ and $\{g_n - r_n\}$ are both convergent in an appropriate sense (we have already seen a similar assertion in Proposition 5.3). The second lemma is also a convergence result, in this case for convex combinations.

Lemma 7.7 *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be probability spaces. Let $\varphi_n : \Omega_1 \to \mathbb{R}$ and $\psi_n : \Omega_2 \to \mathbb{R}$ be measurable, and define $\xi_n : \Omega_1 \times \Omega_2 \to \mathbb{R}$ by*

$$\xi_n(\omega_1, \omega_2) = \varphi_n(\omega_1) + \psi_n(\omega_2).$$

Suppose that $\xi_n \to \xi$ in $\mu_1 \otimes \mu_2$-probability where $\xi : \Omega_1 \times \Omega_2 \to \mathbb{R} \cup \{-\infty\}$. It is clear that we cannot expect φ_n and ψ_n to converge. However, we can find a sequence of real

numbers r_n and measurable functions $\varphi : \Omega_1 \to \mathbb{R} \cup \{-\infty\}$, $\psi : \Omega_2 \to \mathbb{R} \cup \{-\infty\}$ such that $\varphi_{n_k} + r_{n_k}$ converges in probability to φ, $\psi_{n_k} - r_{n_k}$ converge in probability to ψ, and

$$\xi(\omega_1, \omega_2) = \varphi(\omega_1) + \psi(\omega_2) \quad \mu_1 \otimes \mu_2\text{-almost surely,}$$

for some subsequence $\{n_k\}$.

Proof. See [13, Lemma 1].

Lemma 7.8 *Let $f_n : (\Omega, \mathcal{F}, P) \to \mathbb{R}_+ \cup \{\infty\}$ be a sequence of nonnegative random variables, then there exists a sequence $g_n \in conv(f_n, f_{n+1}, \dots)$ that converges almost surely to a function g, where $conv(A) = \{\sum_{i=1}^n t_i a_i : t_i \geq 0, \sum_{i=1}^n t_i = 1, a_i \in A\}$.*

Proof. See the beautiful proof of Delbaen and Schachermayer [5, Lemma A1.1].
Using these two lemmas, we prove

Proposition 7.9 *Let X and Y be Polish spaces, $c : X \times Y \to \mathbb{R}_+ \cup \{\infty\}$ lower semicontinuous. Let $\mu \in P(X), \nu \in P(Y)$ such that c is $\mu \otimes \nu$-almost surely finite, and $\pi \in \Pi(\mu, \nu)$ an optimizer of the Kantorovich problem with $I(\pi)$ finite. Then π is strongly c-monotone.*

Proof. We need to find Borel functions $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ and $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ with $\varphi(x) + \psi(y) \leq c(x, y)$ for any x and any y, and such that equality holds π-almost surely. Let $(\varphi_n, \psi_n) \in \Phi_c$ be a minimizing sequence of bounded functions for J, and $\xi_n(x, y) = \varphi_n(x) + \psi_n(y)$, then by Proposition 1.1,

$$\int_{X \times Y} \xi_n d\pi = J(\varphi_n, \psi_n) = \int_{X \times Y} \xi_n d(\mu \otimes \nu).$$

By the Kantorovich duality, these expressions converge to $I(\pi)$, and without loss of generality the convergence is monotone. The functions $f_n = c - \xi_n$ are nonnegative and have a sequence of convex combinations that converges $\mu \otimes \nu$-almost surely, by Lemma 7.8. Therefore a convex combination of ξ_n converges $\mu \otimes \nu$-almost surely. By possibly replacing φ_n and ψ_n with some convex combination, we may assume without loss of generality that ξ_n converges to ξ $\mu \otimes \nu$-almost surely (the convex combinations are also bounded). Since c is finite $\mu \otimes \nu$-almost surely and $\xi_n \leq c$, ξ is $\mu \otimes \nu$-almost surely finite.
By Lemma 7.7, there exists a sequence $r_n \in \mathbb{R}$ such that almost surely with respect to μ and ν, $\varphi_{n_k} + r_{n_k} \to \varphi$ and $\psi_{n_k} - r_{n_k} \to \psi$ with $\varphi(x) + \psi(y) = \xi(x, y)$. Without modifying the value of $J(\varphi, \psi)$, we can redefine φ and ψ on μ and ν-measure zero sets to have for *any* x, y

$$\xi(x, y) = \varphi(x) + \psi(y) \leq c(x, y).$$

By Fatou's lemma

$$I(c - \xi) \leq \liminf_{n \to \infty} I(c - \xi_n) = 0,$$

and as $c \geq \xi$, equality must hold. This implies $c = \xi = \varphi + \psi$ π-almost surely. The Borel functions φ and ψ are found, and π is strongly c-monotone.
Since any strongly c-monotone set is c-monotone, Proposition 7.9 is a stronger result than Proposition 7.2. In addition, the proof of the latter requires continuity of the cost function while the above proof does not. On the other hand, the proof of Proposition 7.9 is considerably more technical; see the proof of Lemma 7.7.
Before showing that a strongly c-monotone transference plan is optimal, we need one last elementary lemma.

Lemma 7.10 *Let $a, b \in \mathbb{R}$ and for $n \geq 0$ define*

$$a_n \stackrel{\text{def}}{=} \min\left(n, \max(-n, a)\right), \qquad b_n \stackrel{\text{def}}{=} \min\left(n, \max(-n, b)\right), \qquad \xi_n = a_n + b_n.$$

Then $\xi_0 = 0$ and $\{\xi_n\}$ is a monotone sequence that converges to $a + b$.

Proof. This is an excellent example for a lemma that is easier to prove oneself than to read somebody else's proof.

Theorem 7.11 *Let $c : X \times Y \to \mathbb{R}_+ \cup \{\infty\}$ be a lower semi-continuous function on the Polish spaces X and Y. Suppose that $\mu \in P(X)$, $\nu \in P(Y)$ and $\pi \in \Pi(\mu, \nu)$ is strongly c-monotone. Then π is an optimal transference plan between μ and ν.*

Proof. If the value of the Kantorovich problem is infinite, the theorem holds trivially. Otherwise, let π_0 be an optimal transference plan with a finite value for I. Our goal is to show that $I(\pi) \leq I(\pi_0)$. We know that there exist Borel functions such that for any x and any y, $\varphi(x) + \psi(y) \leq c(x, y)$ with equality holding π-almost anywhere. We will use Lemma 7.10 to obtain monotone convergence of bounded functions and then conclude by the monotone convergence theorem. To apply the lemma, define

$$\varphi_n(x) = \min\left(n, \max(-n, \varphi(x))\right), \qquad \psi_n(x) = \min\left(n, \max(-n, \psi(a))\right),$$

and

$$\xi_n(x, y) = \varphi_n(x) + \psi_n(y).$$

Then $\xi_n(x, y) \to \xi(x, y) = \varphi(x) + \psi(y)$ for any x and any y and the convergence is monotone; $\xi_n \nearrow \xi$ on $\{\xi \geq 0\}$ and $\xi_n \searrow \xi$ on $\{\xi \leq 0\}$. Since $\xi = c$ π-almost surely, we have

$$\int_{X \times Y} \xi d\pi = \int_{X \times Y} c d\pi \in [0, \infty].$$

Furthermore, as $\int_{X \times Y} \xi d\pi_0 \leq \int_{X \times Y} c d\pi_0$ and the right hand side is finite, so the left hand side is well-defined and is either finite or equals minus infinity. If φ and ψ are integrable, applying Proposition 1.1 gives

$$\int_{X \times Y} \xi d\pi = \int_{X \times Y} \xi d\pi_0, \tag{7.9}$$

since both sides equal $J(\varphi, \psi)$. The optimality of π follows, as $\xi = c$ π-almost surely and $\xi \leq c$ surely. However, φ and ψ are not a-priori integrable. On the other hand, the functions φ_n and ψ_n are integrable, so for any n,

$$\int_{X\times Y} \xi_n d\pi = \int_X \varphi_n d\mu + \int_Y \varphi_n d\nu = \int_{X\times Y} \xi_n d\pi_0.$$

We express each hand side as the sum of two monotone sequences of functions

$$\int_{X\times Y} \xi_n 1_{\{\xi\geq 0\}} d\pi + \int_{X\times Y} \xi_n 1_{\{\xi<0\}} d\pi = \int_{X\times Y} \xi_n 1_{\{\xi\geq 0\}} d\pi_0 + \int_{X\times Y} \xi_n 1_{\{\xi<0\}} d\pi_0. \quad (7.10)$$

Ideally, we would apply the monotone convergence theorem to each of the four elements separately; this would yield (7.9). However, we need to verify that expressions of the form $\infty - \infty$ do not appear. We proceed by showing that one term in each hand side has a finite limit. Since $\xi = c \geq 0$ π-almost surely, $\int_{X\times Y} \xi_n 1_{\{\xi<0\}} d\pi = 0$ for any n. Therefore, the left hand side of (7.10) converges to

$$\int_{X\times Y} \xi 1_{\{\xi\geq 0\}} d\pi = \int_{X\times Y} \xi d\pi,$$

be it finite or not. On the other hand, $\int_{X\times Y} \xi 1_{\{\xi\geq 0\}} d\pi_0 \leq \int_{X\times Y} c d\pi_0 < \infty$, by the hypothesis on π_0. Therefore, the left integral on the right hand side of (7.10) has a finite limit. Hence the use of the monotone convergence theorem in (7.10) is valid. It results in (7.9), and the theorem follows.

Theorem 7.11 together with Proposition 7.6 imply the converse of Proposition 7.2 when c is finitely valued and lower semi-continuous. Indeed, if π is c-monotone, then it is strongly c-monotone by Proposition 7.6. By Theorem 7.11, π is an optimal transference plan.

7.3 Stability under weak convergence

We conclude this section by proving that when c is continuous, the optimal coupling depends continuously on the marginal measures μ and ν. The idea is to replace optimality by the equivalent c-monotonicity, which behaves well under weak convergence. Now c has to be continuous, because the argument involves approximating points in the support of the limit coupling by points in the support of the couplings in a sequence.

Theorem 7.12 *Let X and Y be Polish spaces and $c : X \times Y \to \mathbb{R}_+$ a finitely valued continuous function. Let $\{\mu_n\}$ and $\{\nu_n\}$ be sequences of probability measures on X and Y respectively that weakly converge to μ and ν. By Proposition 2.2, for any n there exists $\pi_n \in \Pi(\mu_n, \nu_n)$ such that π_n is the optimal coupling of μ_n and ν_n with respect to c; we assume that $I(\pi_n)$ is finite for any n. Then there exists a subsequence π_{n_k} that converges weakly to an optimal coupling of μ and ν. Any (partial) limit of a subsequence of $\{\pi_n\}$ is also an optimizer.*

Proof. An analogous argument to the proof of Proposition 2.1 shows that $\Pi(\mu,\nu)\cup_n \Pi(\mu_n,\nu_n)$ is tight; by Prokhorov's theorem ([2, Theorem 5.1]), $\{\pi_n\}$ has a convergent subsequence π_{n_k}. Therefore, we only have to prove the last statement of the theorem. Let π be a limit point of $\{\pi_n\}$, then there exists a subsequence $\pi_{n_k} \xrightarrow{\mathrm{D}} \pi$. We wish to show that π is c-monotone. By Proposition 7.9, π_{n_k} is strongly c-monotone, thus also c-monotone (alternatively, one can use Proposition 7.2 directly). Since c is continuous, π is also c-monotone: as a consequence of the Portmanteau theorem ([2, Theorem 2.1]), each point $(x,y)\in supp(\pi)$ is a limit of $(x_k,y_k)\in supp(\pi_{n_k})$. Given $(x_i,y_i)\in supp(\pi)$, $i=1,\dots,m$, we have $(x_i^k,y_i^k)\to(x_i,y_i)$ with $(x_i^k,y_i^k)\in supp(\pi_{n_k})$. By continuity of c and c-monotonicity of π_{n_k},

$$\sum_{i=1}^m c(x_i,y_i)=\lim_{k\to\infty}\sum_{i=1}^m c(x_i^k,y_i^k)\le\lim_{k\to\infty}\sum_{i=1}^m c(x_i^k,y_{i-1}^k)=\sum_{i=1}^m c(x_i,y_{i-1}).$$

As π is c-monotone, it is strongly c-monotone by Proposition 7.6, and it is optimal by Theorem 7.11.

Example 1: consider arbitrary measures μ and ν on X and Y respectively. By the strong law of large numbers, one can find discrete probability measures

$$\mu_n=\frac{1}{n}\sum_{i=1}^n\delta\{x_i\},\qquad\qquad \nu_n=\frac{1}{n}\sum_{j=1}^n\delta\{y_j\},$$

such that $\mu_n\xrightarrow{\mathrm{D}}\mu$ and $\nu_n\xrightarrow{\mathrm{D}}\nu$. For any n, an optimal coupling $\pi_n\in\Pi(\mu_n,\nu_n)$ can be found by e.g. the Hungarian method (see Section 4). When c is finitely valued, Theorem 7.12 applies. Then π_n converge, up to a subsequence, to an optimal transference plan μ and ν. Furthermore, by the Portmanteau theorem $I(\pi_n)$ converges to the value of the Kantorovich problem of μ and ν (which is possibly infinite). This is immediate if c is bounded, otherwise one can use truncation. If π is the unique optimal coupling between μ and ν, then any sequence of optimal couplings $\pi_n\in\Pi(\mu_n,\nu_n)$ converges weakly to π. Thus a solution to the Monge–Kantorovich problem can be approximated by a discrete measure, which is a solution to a discrete Monge–Kantorovich problem.

When uniqueness fails to hold, it is easy to construct sequences $\mu_k\to\mu$ and $\nu_k\to\nu$ such that the optimal couplings $\pi_k\in\Pi(\mu_k,\nu_k)$ do not converge. We conclude this section by such example.

Example 2: Let $x_1=(0,1)$, $x_2=(1,0)$, $y_1=(1,1)$, $y_2=(0,0)$, $\mu=(\delta\{x_1\}+\delta\{x_2\})/2$ and $\nu=(\delta\{y_1\}+\delta\{y_2\})/2$ with quadratic cost (see Figure 3). The two permutations $\sigma_1=id$ and $\sigma_2=(12)$. Both induce a cost of $(1+1)/2=1$ and are therefore optimal. We can perturb the measure μ by an arbitrarily small amount so that any of the two permutations is optimal: for $1/2>\varepsilon>0$ take

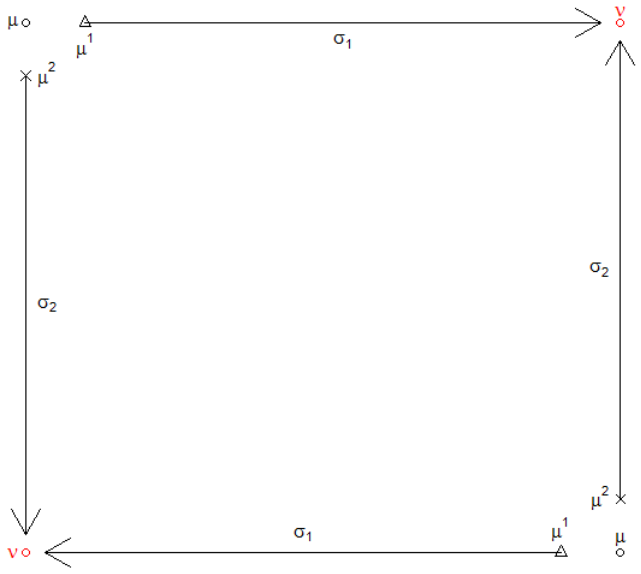
$$w_1=(\varepsilon,1),\quad w_2=(1-\varepsilon,0),\qquad \mu^1=\frac{\delta\{w_1\}+\delta\{w_2\}}{2}$$

and

$$z_1 = (0, 1-\varepsilon), \quad z_2 = (1, \varepsilon), \qquad \mu^2 = \frac{\delta\{z_1\} + \delta\{z_2\}}{2}.$$

For any $\varepsilon > 0$, σ_1 is the unique optimal map from μ^1 to ν and σ_2 is uniquely optimal

Figure 3: Two perturbations μ^1, μ^2 of the source measure μ. The target measure is ν, and σ_i is the optimal transport map from μ^1 to ν



between μ^2 and ν. In other words, the optimal couplings are

$$\pi^1 = \frac{\delta\{(w_1, y_1)\} + \delta\{(w_2, y_2)\}}{2} \quad \text{and} \quad \pi^2 = \frac{\delta\{(z_2, y_1)\} + \delta\{(z_1, y_2)\}}{2}.$$

Clearly $\pi^1 \xrightarrow{\mathrm{D}} (\delta\{(x_1, y_1)\} + \delta\{(x_2, y_2)\})/2$ and $\pi^2 \xrightarrow{\mathrm{D}} (\delta\{(x_1, y_2)\} + \delta\{(x_2, y_1)\})/2$ so they are far from each other; but $(\mu^1 - \mu^2) \xrightarrow{\mathrm{D}} 0$ when $\varepsilon \to 0$.

Acknowledgments

I would like to thank Victor Panaretos for introducing me to the beautiful field of optimal transportation. The discussions with him, his suggestions and his comments improved not only this text, but also my understanding of this fascinating topic.

I also thank Christian Mazza for his willingness to undertake the task of being the external examinator for this project.

Several people were kind enough to proof-read parts of this text. Thanks go to Shahin Tavakoli and in particular Amos Zemel, whose suggestions greatly ameliorated the level of the English of this project.

Mikael Kuusela has given me useful tips for the typesetting; I thank him for that.

Last, but not least, I would like to express my deep gratitude to the *Commission fédérale des bourses pour étudiants étrangers* of the Swiss government for financing my studies by a scholarship.

References

[1] L. Ambrosio and A. Pratelli. Existence and stability results in the l^1–theory of optimal transportation. 1813:123–160, 2003.

[2] P. Billingsley. *Convergence of probability measures*, volume 137. second ed. John Wiley&Sons Inc., New York, 1999.

[3] R. G. Bland. New finite pivoting rules for the simplex method. *Mathematics of operations research*, 2:103–107, 1977.

[4] R. Chartrand, B. Wohlberg, K. R. Vixie, and E. M. Bollt. A gradient descent solution to the Monge–Kantorovich problem. *Applied Mathematical Sciences*, 3:1071–1080, 2009.

[5] F. Delbaen and W. Schachermayer. A general version of the fundamental theorem of asset pricing. *Mathematische Annalen*, 300:463–520, 1994.

[6] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Math.*, 177:113–161, 1996.

[7] L. V. Kantorovich. On the translocation of masses. *(Dokl.) Acad. Sci. URSS 37*, 3:199–201, 1942.

[8] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2:83–97, 1955.

[9] D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming*. Springer, 2008.

[10] G. Monge. Mémoire sur la théorie des déblais et de remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 177:666–704, 1781.

[11] J. Munkers. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5:32–38, 1957.

[12] R. T. Rockafellar. *Convex analysis*, volume Reprint of the 1970 original, Princeton Paperbacks. Princeton University Press, Princeton, NJ, 1997.

[13] W. Schachermayer and J. Teichmann. Characterization of optimal transport plans for the Monge–Kantorovich problem. *Proceedings of the American Mathematical Society*, 137:519–529, 2009.

[14] C. Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Society, 2003.

[15] J.-C. Wei. Lecture notes for linear programming. `http://www.math.cuhk.edu.hk/~wei/lpch4.pdf`. Retrieved June 8, 2012.