January 4, 2001

Lectures on Geometry

Christer O. Kiselman

Contents:

1. Introduction

These notes comprise the main part of a course presented at Uppsala University in the Fall of 1999. (The original name of the course was *Geometry and Analysis.*) My initial idea was to present duality in geometry, or how to use analysis to solve geometric problems. This I did, not dodging the more difficult aspects of this duality—see Theorems 5.10 and 5.11 and the nasty Examples 5.7–5.9 and 5.12. A natural application was Kantorovich's theorem, to which I found a new proof (Theorem 6.1). Also duality of sets which are not convex was considered (Section 8). Finally, my interest in image analysis led me to an attempt to survey some results in digital geometry; since existing proofs of Khalimsky's Jordan curve theorem turned out to be difficult to present, I tried to find a simple one, which appears now in Section 11.

I am grateful to Björn Ivarsson for valuable criticism of an early draft of Section 2; to Thomas Kaijser for sharing his knowledge on the Monge–Kantorovich problem; and to Ingela Nyström and Erik Palmgren for helpful comments on an early version of Section 11.

2. Closure operators and Galois correspondences

An *order relation* in a set X is a relation (a subset of X^2) which satisfies three conditions: it is *reflexive*, *antisymmetric* and *transitive*. This means, if we denote the relation by $\leqslant$, that for all $x, y, z \in X$,

$$x \leqslant x; \tag{2.1}$$

$$x \leqslant y \text{ and } y \leqslant x \text{ implies } x = y; \tag{2.2}$$

$$x \leqslant y \text{ and } y \leqslant z \text{ implies } x \leqslant z. \tag{2.3}$$

An *ordered set* is a set X together with an order relation. (Sometimes one says *partially ordered set.*)

A basic example is the set $P(W)$ of all subsets of a set W, with the order relation given by inclusion, thus $A \leqslant B$ being defined as $A \subset B$ for $A, B \in P(W)$.

A *closure operator* in an ordered set X is a mapping $X \ni x \mapsto \overline{x} \in X$ which is *expanding, increasing* (or *order preserving*), and *idempotent*; in other words, which satisfies the following three conditions for all $x, y \in X$:

$$x \leqslant \overline{x}; \tag{2.4}$$

$$x \leqslant y \text{ implies } \overline{x} \leqslant \overline{y}; \tag{2.5}$$

$$\overline{\overline{x}} = \overline{x}. \tag{2.6}$$

In checking (2.6) it is of course enough to prove that $\overline{\overline{x}} \leqslant \overline{x}$ if we have already proved (2.4).

The element $\overline{x}$ is said to be the *closure* of x. Elements x such that $\overline{x} = x$ are called *closed* (for this operator). An element is closed if and only if it is the closure of some element (and then it is the closure of itself).

A basic example of a closure operator is of course the topological closure operator which associates to a set in a topological space its topological closure, i.e., the smallest closed set containing the given set. In fact a closure operator in $P(W)$ defines a topology in W if and only if it satisfies, in addition to (2.4), (2.5), (2.6) above, two extra conditions, viz. that $\overline{\varnothing} = \varnothing$ and

$$\overline{A \cup B} = \overline{A} \cup \overline{B} \text{ for all } A, B \subset W. \tag{2.7}$$

Another closure operator of great importance is the operator which associates to a set in $\mathbf{R}^n$ its convex hull, the smallest convex set containing the given set.

In both these examples X is the power set of some set W, and the closure operator is given as an intersection:

$$\overline{A} = \bigcap (Y; Y \text{ is closed and } Y \supset A).$$

More generally, if a closure operator $x \mapsto \overline{x}$ is given and we denote by F the set of its closed elements, then

$$\overline{x} = \inf(y \in F; y \geqslant x). \tag{2.8}$$

Conversely, any subset F of X such that the infimum of a subset of F always exists defines a closure operator by formula (2.8).

A *Galois correspondence* is a pair (f, g) of two decreasing mappings $f: X \to Y$, $g: Y \to X$ of two given ordered sets X, Y such that $g \circ f$ and $f \circ g$ are expanding. In other words we have $f(x_1) \geqslant f(x_2)$ and $g(y_1) \geqslant g(y_2)$ if $x_1 \leqslant x_2$ and $y_1 \leqslant y_2$, and $g(f(x)) \geqslant x$ and $f(g(y)) \geqslant y$ for all $x \in X$, $y \in Y$; Kuroš [1962:6:11].

The name Galois correspondence alludes to the first correspondence of that kind, established by Galois[1] with X as the subsets of a field, Y as the sets of isomorphisms of this field, $f(x)$ as the group of all isomorphisms leaving all elements of x invariant, and $g(y)$ as the subfield the elements of which are left fixed by all elements of y.

[1] Évariste Galois, 1811–1832.

Proposition 2.1. *Let $f\colon X \to Y$, $g\colon Y \to X$ be a Galois correspondence. Then $g \circ f\colon X \to X$ and $f \circ g\colon Y \to Y$ are closure operators. Moreover, $f \circ g \circ f = f$ and $g \circ f \circ g = g$.*

Proof. That $g \circ f$ and $f \circ g$ are expanding is part of the definition of a Galois correspondence; that they are increasing follows from the fact that they are compositions of two decreasing mappings. We know that $f \circ g$ is expanding, so $(f \circ g)(f(x)) \geqslant f(x)$; thus $f \circ g \circ f \geqslant f$. On the other hand, also $g \circ f$ is expanding, i.e., $g \circ f \geqslant \mathrm{id}_X$, so $f \circ g \circ f \leqslant f \circ \mathrm{id}_X = f$, hence $f \circ g \circ f = f$. By symmetry, $g \circ f \circ g = g$. From either one of these identities we easily obtain that $g \circ f$ and $f \circ g$ are idempotent.

It is now natural to ask whether the closure operators one obtains from Galois correspondences have some special property. The answer is no: every closure operator comes in a trivial way from some Galois correspondence.

Proposition 2.2. *Let $x \mapsto \overline{x}$ be a closure operator defined in an ordered set X. Then there exist an ordered set Y and a Galois correspondence $f\colon X \to Y$, $g\colon Y \to X$ such that $\overline{x} = g(f(x))$ for all $x \in X$.*

Proof. We define Y as the set of all closed elements in X with the opposite order, thus $y_1 \leqslant_Y y_2$ shall mean that $y_1 \geqslant_X y_2$. Let $f\colon X \to Y$ and $g\colon Y \to X$ be defined by $f(x) = \overline{x}$ and $g(y) = y$. Then both f and g are decreasing, and $g \circ f(x) = \overline{x} \geqslant_X x$, $f \circ g(y) = y \geqslant_Y y$. So $g \circ f$ and $f \circ g$ are expanding, and $\overline{x} = g(f(x))$ as desired.

This proposition is, in a sense, completely uninteresting. This is because the Galois correspondence is obtained from X and the closure operator in a totally trivial way. However, there are many Galois correspondences in mathematics that are highly interesting and represent a given closure operator. This is because they allow for important calculations to be made or for new insights into the theory.

We now ask whether the composition of two closure operators is a closure operator.

Example. Let A be the set of all points (x, y) in $\mathbf{R}^2$ satisfying $y \geqslant 1/(1 + x^2)$. This is a closed set, so $\overline{A} = A$ if we let the bar denote topological closure. The convex hull of A is the set $\operatorname{cvx} A = \{(x, y) \in \mathbf{R}^2;\ y > 0\}$, which is not closed; its closure is $\overline{\operatorname{cvx} A} = \{(x, y);\ y \geqslant 0\}$. Hence we see that it is not true that the convex hull of a closed set is closed; on the other hand we shall see that the closure of a convex set is convex. Define $f(A) = \operatorname{cvx} A$ and $g(A) = \overline{A}$. Then the composition $h = g \circ f$ is a closure operator, whereas the other composition $k(A) = f \circ g(A) = \operatorname{cvx} \overline{A}$ defines a mapping $k = f \circ g$ which is not a closure operator ($k \circ k \neq k$).

Proposition 2.3. *Let $f, g\colon X \to X$ be two closure operators. The following properties are equivalent:*

(i) *$g \circ f$ is a closure operator;*
(ii) *$g \circ f$ is idempotent;*
(iii) *$g \circ f \circ g = g \circ f$;*
(iv) *$f \circ g \circ f = g \circ f$;*
(v) *$g(x)$ is f-closed if x is f-closed.*

If one of these conditions is satisfied, then $g \circ f$ is the supremum of the two closure operators f and g in the ordered set of all closure operators; moreover $f \circ g \leqslant g \circ f$.

Proof. That $h = g \circ f$ is expanding and increasing is true for any composition of expanding and increasing mappings, so it is clear that (i) and (ii) are equivalent. It is also easy to see that (iv) and (v) are equivalent. If (iii) holds, then $h \circ h = g \circ f \circ g \circ f = g \circ f \circ f = g \circ f = h$, so h is idempotent. Similarly, (iv) implies (ii). Conversely, if (ii) holds, then

$$g \circ f \leqslant g \circ f \circ g \leqslant h \circ h = h = g \circ f$$

and

$$g \circ f \leqslant f \circ g \circ f \leqslant h \circ h = h = g \circ f,$$

so we must have equality all the way in both chains of inequalities, which proves that (iii) and (iv) hold. The last statement is easy to verify.

Corollary 2.4. *Two closure operators f and g commute if and only if both $g \circ f$ and $f \circ g$ are closure operators.*

Proof. If $g \circ f = f \circ g$, then (iii) obviously holds, so $g \circ f$ is a closure operator. Conversely, if $g \circ f$ is a closure operator, then (iii) applied to f and g says that $g \circ f \circ g = g \circ f$; if also $f \circ g$ is a closure operator, then (iv) applied to g and f says that $g \circ f \circ g = f \circ g$. Thus f and g commute.

When two closure operators f and g are given, it may happen that $f \leqslant g$. Then the semigroup generated by f and g consists of at most three elements, id_X, f, g. If both $g \circ f$ and $f \circ g$ are closure operators, then the semigroup generated by f and g has at most four elements, id_X, f, g, and $g \circ f = f \circ g$. If precisely one of $g \circ f$ and $f \circ g$ is a closure operator, then the semigroup generated has exactly five elements, $\mathrm{id}_X, f, g, g \circ f$, and $f \circ g$, of which four are closure operators. When none of $g \circ f$ and $f \circ g$ is a closure operator, the semigroup of all compositions $f_m \circ \cdots \circ f_1$, with $f_j = f$ or $f_j = g$, $m \in \mathbf{N}$, may be finite or infinite.

Applying Proposition 2.3 to the case $f(A) = \operatorname{cvx} A$, $g(A) = \overline{A}$ we see that the operation of taking the topological closure of the convex hull, $A \mapsto \overline{\operatorname{cvx} A}$ is a closure operator. We call $\overline{\operatorname{cvx} A}$ the *closed convex hull* of A. This is a case where the semigroup generated by f and g consists of five elements.

Example. Let E be a finite-dimensional vector space over $\mathbf{R}$ and let E^* denote its dual. (We can think of $\mathbf{R}^n$, but it is often clarifying to distinguish E and its dual.) We shall define a Galois correspondence: the ordered set X shall be the power set of E, the set of all subsets of E, with inclusion as the order, and the set $Y = [-\infty, +\infty]^{E^*}$ shall be the set of all functions with values in the extended real line and defined on the dual of E, the order being the opposite of the usual order, defined by set inclusion of the epigraphs, so that $\varphi \leqslant_Y \psi$ iff $\varphi(\xi) \geqslant \psi(\xi)$ for all $\xi \in E^*$; see Definition 3.10. Thus the constant $+\infty$ is the smallest function, corresponding to vacuum, while $-\infty$ is the largest element, corresponding to an infinitely dense neutron star. We define mappings $f\colon X \to Y$ and $g\colon Y \to X$ as follows. Let H_A denote the *supporting function* of a subset A of E, i.e.,

$$H_A(\xi) = \sup_{x \in A} \xi(x), \qquad \xi \in E^*.$$

Let $f(A) = H_A$. To any function φ on E^* with values in the extended real line $[-\infty, +\infty] = \mathbf{R} \cup \{-\infty, +\infty\}$ we associate a set

$$g(\varphi) = \bigcap_{\xi \in E^*} \{x;\ \xi(x) \leqslant \varphi(\xi)\}.$$

It is an intersection of closed half-spaces (including possibly the whole space and the empty set). Then both f and g are decreasing, and $g \circ f$ and $f \circ g$ are expanding as shown by the formulas

$$(g \circ f)(A) = \bigcap_{\xi \in E^*} \{x \in E;\ \xi(x) \leqslant H_A(\xi)\} \supset A \quad \text{for all sets} A,$$

$$(f \circ g)(\varphi) = H_{g(\varphi)} \geqslant_Y \varphi \quad \text{for all functions } \varphi.$$

This is a highly interesting Galois correspondence. We shall determine its closed elements in Section 5.

3. Convex sets and functions

Given two points a, b in a vector space E we define the *segment* between a and b as the set

$$[a, b] = \{(1-t)a + tb;\ 0 \leqslant t \leqslant 1\}.$$

A subset A of E is said to be *convex* if it contains the whole segment $[a, b]$ as soon it contains a and b. The convex subsets of the real line are precisely the intervals. The definition can therefore be given as follows: for any affine mapping $\varphi\colon \mathbf{R} \to E$ the inverse image $\varphi^{-1}(A)$ shall be an interval.

The notion of a convex set has a sense in affine spaces, which, roughly speaking, are like vector spaces without a determined origin. In affine spaces sums like $\sum \lambda_j x_j$ have a sense if and only if $\sum \lambda_j = 1$, which is the case in the sums we need to work with when considering convexity. If under this hypothesis we perform a translation by a vector a, then form the sum, and finally perform a translation by the vector $-a$, we get a result independent of the translation. Indeed, $\sum \lambda_j (x_j - a) + a = \sum \lambda_j x_j$.

The *convex hull* of a set X is the set

(3.1) $$\operatorname{cvx} X = \bigcap_Y (Y; Y \text{ is convex and contains } X)\,;$$

cf. (2.8). It is clear that $\operatorname{cvx} X$ is convex and that it is the smallest convex set containing X.

The description (3.1) of the convex hull is a description from the outside: we approach the hull by convex sets containing it. There is also a description from within:

Theorem 3.1. *Let X be any subset of a vector space E. Then $\operatorname{cvx} X$ is the set of all linear combinations $\sum_{j=1}^N \lambda_j x_j$, where N is an arbitrary integer $\geqslant 1$, where the points x_j belong to X, and where the coefficients λ_j satisfy $\lambda_j \geqslant 0$, $\sum \lambda_j = 1$.*

The proof of this proposition is left to the reader; cf. Hiriart-Urruty & Lemaréchal [1993: Prop. 1.3.4], Kiselman [1991a or b: Theorem 2.1], or Rockafellar [1970: Theorem 2.3].

The linear combinations that occur in the proposition are called *convex combinations.* (The requirement that N be at least 1 is important; thanks to this requirement we obtain that the convex hull of the empty set is empty.)

Theorem 3.2 (Carathéodory's[2] theorem). *Let X and E be as in Theorem 3.1 and assume that E is of finite dimension n. Then it is enough to take N in the representation equal to $n+1$.*

For the proof of this result see Kiselman [1991a or b: Theorem 6.1] or Rockafellar [1970: Theorem 17.1]. In one and two dimensions, maybe even three, it is intuitively obvious.

Corollary 3.3. *The convex hull of a compact subset of $\mathbf{R}^n$ is compact.*

Proof. It is clear that for $N = 1, 2, \ldots$ the set

$$K_N = \Big\{ \textstyle\sum_1^N \lambda_j x_j ; \lambda_j \geqslant 0,\ \sum_1^N \lambda_j = 1 \Big\}$$

is compact, for it is the continuous image of a compact set under a continuous mapping. In general, the convex hull of K is the union of all the K_N, $N \geqslant 1$. But thanks to Carathéodory's theorem, the sequence is stationary and the union of the sequence is equal to K_{n+1}.

Theorem 3.4. *The closure of a convex set in a topological vector space is convex; more generally, this is true in a vector space E equipped with a topology such that all translations $x \mapsto x - a$, $a \in E$, and all dilations $x \mapsto \lambda x$, $\lambda \in \mathbf{R}$, are continuous.*

Proof. Suppose that A is convex and let $a_0, a_1 \in \overline{A}$. We have to prove that $[a_0, a_1] \subset \overline{A}$. Consider $a = (1-t)a_0 + ta_1$ for an arbitrary fixed $t \in [0,1]$, and let V be an arbitrary neighborhood of a. Then we can choose a neighborhood V_0 of a_0 such that $(1-t)V_0 + ta_1 \subset V$. Since a_0 belongs to the closure of A, there exists a point $b_0 \in A \cap V_0$. Next we can find a neighborhood V_1 of a_1 such that $(1-t)b_0 + tV_1 \subset V$. There exists a point $b_1 \in A \cap V_1$. Hence the linear combination $(1-t)b_0 + tb_1$ belongs to A and since it also belongs to V, we have proved that V intersects A, thus that $a \in \overline{A}$.

Thanks to this theorem we can state that $\overline{\operatorname{cvx} X}$ is the smallest closed and convex set which contains X. The set is called the *closed convex hull* of X.

We shall now study the possibility of separating two convex sets by a hyperplane.

A *hyperplane* is an affine subspace of codimension one; in other words it consists of all solutions to a single linear equation $\xi(x) = b$, where ξ is a nonzero linear form on the space and where b is a real number. In general we need to distinguish between a general hyperplane, defined by a not necessarily continuous linear form on the one hand, and a closed hyperplane, defined by a continuous linear form on the other hand. However, we shall restrict attention here to finite-dimensional spaces equipped with the unique separated vector space topology—when in the sequel we consider a finite-dimensional vector space we shall always assume that it carries this topology. For these finite-dimensional spaces all linear forms are continuous and all affine subspaces closed. To every hyperplane we associate two closed half-spaces, $D_+ = \{x;\ \xi(x) \geqslant b\}$ and $D_- = \{x;\ \xi(x) \leqslant b\}$. There is a choice to be made here, since $-\xi(x) = -b$ defines the same hyperplane as $\xi(x) = b$.

We shall say that a hyperplane $H = \{x;\ \xi(x) = b\}$ *separates* two subsets X and Y of E if $\xi(x) \leqslant b \leqslant \xi(y)$ for all $x \in X$ and all $y \in Y$. We shall say that H *separates* X and Y *strictly* if we have $\xi(x) < b < \xi(y)$ for all $x \in X$ and all $y \in Y$.

[2] Constantin Carathéodory, 1873–1950.

Theorem 3.5. *Let F be a closed convex subset of a finite-dimensional vector space E, and $y \in E$ a point which does not belong to F Then there exists a hyperplane which separates F and y strictly, even in the strong sense that for some numbers b_0, b_1 we have $\xi(x) < b_0 < b < b_1 < \xi(y)$ for all $x \in F$.*

Theorem 3.6. *Let A be a convex subset of a finite-dimensional vector space E, and $y \in E$ a point which does not belong to A or which belongs to the boundary of A. Then there exists a hyperplane which separates A and y.*

Theorem 3.7. *Let F and K be closed convex subsets of a finite-dimensional vector space E, and assume that they are disjoint and that K is compact. Then there exists a hyperplane which separates F and K strictly, even in the strong sense that for some numbers b_0, b_1 we have $\xi(x) < b_0 < b < b_1 < \xi(y)$ for all $x \in F$ and all $y \in K$.*

Theorem 3.8. *Let X and Y be convex subsets of a finite-dimensional vector space E, and assume that they are disjoint. Then there exists a hyperplane which separates X and Y.*

Proof of Theorem 3.5. We may assume that $y = 0$. If $F = \emptyset$ the result is certainly true, so we may assume that F is nonempty. Let $d = \inf_{x \in F} \|x\|$, where we use a Euclidean norm. Then d is positive, since F is closed, and $d < +\infty$, since F is nonempty. Moreover, there exists a point a in (the possibly unbounded set) F where the infimum is attained. This is because the infimum over all of F is the same as the infimum over the compact set $K = \{x \in F; \|x\| \leqslant 2d\}$. I claim that $a \cdot x \geqslant \|a\|^2$ for all $x \in F$. Study the function $f(t) = \|(1-t)a + tx\|^2$, where x is a point in F. We must have $f'(0) \geqslant 0$, for if $f'(0)$ were negative, $f(t)$ would be smaller than $f(0) = d^2$ for small positive t contrary to the definition of d and a; note that all points $(1-t)a + tx$ with $t \in [0,1]$ belong to F. Now it is easy to calculate f':

$$f'(t) = -2(1-t)\|a\|^2 - 2ta \cdot x + 2(1-t)a \cdot x + 2t\|x\|^2;$$

in particular $f'(0) = -2\|a\|^2 + 2a \cdot x$. Since this quantity has to be nonnegative, we must have $a \cdot x \geqslant a \cdot a$ as claimed. Any hyperplane $a \cdot x = b$ with $0 < b < a \cdot a$ is now strictly separating.

Proof of Theorem 3.6. Let $(a_j)_{j \in \mathbf{N}}$ be a sequence which is dense in A. We apply Theorem 3.5 to the compact set $F_m = \operatorname{cvx}\{a_0, ..., a_m\}$, $m \in \mathbf{N}$. If y does not belong to A, it does not belong to F_m either. If y belongs to the boundary of A, it may belong to A, and some care is needed to avoid that y belongs to F_m. We can choose $y_m \notin A$ such that y_m tends to y and apply Theorem 3.5 to F_m and y_m.

For every m there is a hyperplane $\xi_m(x) = b_m$ which separates y_m and F_m. We may assume that $\|\xi_m\| = 1$. Then the sequence (b_m) is also bounded, so there are subsequences $(\xi_{m_j})_j$ and $(b_{m_j})_j$ which converge to limits ξ and b respectively. The hyperplane $\xi(x) = b$ separates y from A.

Proof of Theorem 3.7. The set $F - K$ is closed and does not contain the origin. If $\xi(x) = b$ is a hyperplane which separates the origin from $F - K$ in the strong sense indicated in Theorem 3.5, we know that $0 = \xi(0) < b_0 < b < b_1 < \xi(x-y)$ for all $x \in F$ and all $y \in K$, so that

$$b_1 \leqslant \inf_{x \in F, y \in K} (\xi(x-y)) = \inf_{x \in F} \xi(x) - \sup_{y \in K} \xi(y),$$

implying that $\sup_K \xi + b_1 \leqslant \inf_F \xi$, an inequality which proves the theorem.

Proof of Theorem 3.8. We observe that if X and Y are as in the theorem, then $A = X - Y$ does not contain the origin. We apply Theorem 3.6 and argue as in the proof of Theorem 3.7.

An immediate consequence of Theorem 3.5 is that every closed convex set is equal to the intersection of all the closed half-spaces which contain it. But the proof actually gives more: the half-spaces used to form the intersection can all be chosen so that their boundaries contain a point of the given set. More precisely, we considered the largest open ball $B(y, r)$ with center at $y \notin F$ which does not meet F and found that there is a unique point a common to its closure and F; the tangent plane to the ball at a is a separating hyperplane. A *supporting half-space* of a set A is a half-space D such that its boundary H (a hyperplane) contains a point of A. We also say in this situation that H is a *supporting hyperplane* of A. Sometimes, to a given separating hyperplane there is a parallel hyperplane which is supporting, but not always:

Example. The set $\{(x, y) \in \mathbf{R}^2; x > 0, xy \geqslant 1\}$ is convex, and the hyperplane $H = \{(x, y); x = 0\}$ lies in its complement, but there is no supporting hyperplane parallel to H.

Theorem 3.9. *Every closed convex set in a finite-dimensional vector space is equal to an intersection of closed, supporting half-spaces.*

Proof. We just have to note that the half-space $D = \{x; x \cdot a \geqslant a \cdot a\}$ found in the proof of Theorem 3.5 is indeed a supporting half-space: the point a lies both in F and the hyperplane which bounds D.

Definition 3.10. *Let $f: X \to [-\infty, +\infty]$ be a function defined on an arbitrary set X. Then its* ***epigraph*** *is the subset of $X \times \mathbf{R}$ defined as*

$$\operatorname{epi} f = \{(x, t) \in X \times \mathbf{R}; t \geqslant f(x)\}.$$

The ***strict epigraph*** *is the set*

$$\operatorname{epi}_s f = \{(x, t) \in X \times \mathbf{R}; t > f(x)\}.$$

Definition 3.11. *A function $f: X \to [-\infty, +\infty]$ defined on a subset X of a vector space E is said to be* ***convex*** *if its epigraph is a convex set in $E \times \mathbf{R}$.*

It is equivalent to require that the strict epigraph be convex. We shall use the notation $CVX(X)$ for the set of all convex functions defined in a set $X \subset E$ and with values in the extended real line $[-\infty, +\infty]$.

If we extend f to a function g defined in all of E by putting $g(x)$ equal to $+\infty$ outside X we see that the extended function is convex at the same time as f, since the two functions have the same epigraph. Therefore we may always assume (if we like) that convex functions are defined in the whole space.

The constants $+\infty$ and $-\infty$ are convex, since their epigraphs are, respectively, $\emptyset$ and all of E. Furthermore, the maximum $\max(f_1, ..., f_m)$ of finitely many convex functions is convex, since its epigraph is the intersection of the convex sets $\operatorname{epi} f_j$.

However, if we want to prove for instance that the sum of two convex functions is convex, it is convenient to have an inequality to test: the inequality which says that the graph hangs below the chord. Since convex functions may assume both $+\infty$ and $-\infty$ as values, we must handle undefined sums like $(+\infty)+(-\infty)$. It is convenient to introduce *upper addition* $\overset{\cdot}{+}$ and *lower addition* $\underset{\cdot}{+}$. These operations are extensions of the usual addition $\mathbf{R}\times\mathbf{R}\ni(x,y)\mapsto x+y\in\mathbf{R}$. We define without hesitation $(+\infty)+(+\infty)=+\infty$, $(-\infty)+(-\infty)=-\infty$, and use these definitions for both $\overset{\cdot}{+}$ and $\underset{\cdot}{+}$. In the ambiguous cases we define

$$(+\infty)\overset{\cdot}{+}(-\infty)=+\infty,\qquad (+\infty)\underset{\cdot}{+}(-\infty)=-\infty. \tag{3.2}$$

Upper addition defines an upper semicontinuous function $[-\infty,+\infty]^2\to[-\infty,+\infty]$, and, similarly, lower addition a lower semicontinuous mapping.

Very useful rules for computing infima and suprema are

$$\inf_{x\in X}(a\overset{\cdot}{+}f(x))=a\overset{\cdot}{+}\inf_{x\in X}f(x),\qquad \sup_{x\in X}(a\underset{\cdot}{+}f(x))=a\underset{\cdot}{+}\sup_{x\in X}f(x), \tag{3.3}$$

which are valid without exception. The proof of (3.3) consists in checking the equalities for $a=+\infty,-\infty$ and X empty.

Proposition 3.12. *Let E be a vector space. A function $f\colon E\to[-\infty,+\infty]$ is convex if and only if*

$$f\big((1-t)x_0+tx_1\big)\leqslant(1-t)f(x_0)\overset{\cdot}{+}tf(x_1)$$

for all $x_0,x_1\in E$ and all numbers t with $0<t<1$.

We leave the simple proof as an exercise.

The proposition implies that the values of a convex function must satisfy an infinity of inequalities, and in general that means that the values are severely restricted. However, note that this is not always the case. As an example consider a function which is $+\infty$ for $\|x\|>1$ and $-\infty$ for $\|x\|<1$. Such a function is convex irrespective of its values on the unit sphere $\|x\|=1$. This phenomenon is related to the fact that all sets A containing the open unit ball and contained in the closed unit ball are convex. (Of course in these examples we need Euclidean norms, or at least strictly convex norms.) We shall exploit this phenomenon in Section 8.

The *effective domain* of a function $f\colon X\to[-\infty,+\infty]$ is the set where it is smaller than plus infinity:

$$\operatorname{dom} f=\{x\in X;\ f(x)<+\infty\}. \tag{3.4}$$

If f is convex, so is its effective domain.

Theorem 3.13. *Let E be a finite-dimensional vector space, let $f\colon E\to[-\infty,+\infty]$ be a convex function and define $\Omega=(\operatorname{dom} f)^\circ$. (Thus Ω is the interior of the set where the function is less than plus infinity.) Then Ω is convex and the restriction of f to Ω is either equal to the constant $-\infty$ or else a real-valued continuous function.*

We leave the proof as an exercise. See Kiselman [1991a or b, Theorem 9.5].

Corollary 3.14. *Let f and Ω be as in the theorem and suppose that there exists a point $x \in \Omega$ such that $f(x)$ is real. Then there exists a linear functional ξ on E such that $\xi(x) \leqslant f(x)$ for all $x \in E$. Moreover, $g(x) = \liminf_{y\to x} f(y)$ is convex and does not take the value $-\infty$ if $f > -\infty$. We have $g = f$ except on $\partial\Omega$.*

In analogy with (3.1) we define the *convex hull* of a function as the largest convex minorant of the function:

$$\operatorname{cvx} f = \sup(g \in CVX(E); g \leqslant f). \tag{3.5}$$

This is the approach from below. There is also an approach from above, as in Theorem 3.1:

Theorem 3.15. *Let $f\colon E \to [-\infty, +\infty]$ be any function on a vector space E. Then its convex hull is given as an infimum of linear combination of values of f:*

$$\operatorname{cvx} f(x) = \inf\left[\sum_1^N \lambda_j f(x_j); N \geqslant 1, \lambda_j > 0, \sum_1^N \lambda_j = 1, \sum_1^N \lambda_j x_j = x\right], \quad x \in E. \tag{3.6}$$

The proof is analogous to that of Theorem 3.1.

When f is positively homogeneous, i.e., $f(tx) = tf(x)$ for all $t > 0$ and all $x \in E$, then (3.6) can be simplified to

$$\operatorname{cvx} f(x) = \inf\left[\sum_1^N f(x_j); N \geqslant 1, \sum_1^N x_j = x\right], \qquad x \in E. \tag{3.7}$$

4. Infimal convolution

Definition 4.1. *Let G be an abelian group and let $f, g\colon G \to [-\infty, +\infty]$ be two functions defined on G and with values in the extended real axis. Then their infimal convolution $f \mathbin{\square} g$ is defined by*

$$(f \mathbin{\square} g)(x) = \inf_{y \in G}\Big(f(y) \dot{+} g(x-y)\Big), \qquad x \in G. \tag{4.1}$$

Often we take $G = \mathbf{R}^n$, but it is important to note that the definition works in any abelian group G. In image analysis $\mathbf{Z}^2$ and, more generally, $\mathbf{Z}^n$ are very common groups.

Points outside $\operatorname{dom} f$ and $\operatorname{dom} g$ play no role in (4.1). This is in accordance with the interpretation already mentioned of the constant $+\infty$ as vacuum and of $-\infty$ as an infinitely dense neutron star. Think of e^{-f} as a particle—there is an interesting analogy between infimal convolution of f and g and ordinary convolution of e^{-f} and e^{-g}; see Kiselman [1999a or b, Chapters 8 and 9].

Infimal convolution generalizes vector addition (Minkowski addition) of sets. If X and Y are subsets of G we have

$$i_X \mathbin{\square} i_Y = i_{X+Y}, \tag{4.2}$$

where i_X denotes the *indicator function* of the set X; it takes the value 0 in X and $+\infty$ in its complement.

Another important relation to vector addition—a relation which can actually serve to define infimal convolution—is

$$\operatorname{epi}_s(f \mathbin{\square} g) = \operatorname{epi}_s(f) + \operatorname{epi}_s(g), \tag{4.3}$$

where the plus sign denotes vector addition in $G \times \mathbf{R}$. This equation leads to a geometric interpretation of infimal convolution.

For a thorough study of infimal convolution, see Strömberg [1996].

Proposition 4.2. *The infimal convolution is commutative and associative.*

Proof. Since vector addition is commutative and associative, the same is true for infimal convolution in view of (4.3).

Thanks to this proposition we can define generally $f_1 \mathbin{\square} \cdots \mathbin{\square} f_m$; no parantheses are needed. It is easy to see that

$$\operatorname{dom}(f_1 \mathbin{\square} \cdots \mathbin{\square} f_m) = \operatorname{dom} f_1 + \cdots + \operatorname{dom} f_m. \tag{4.4}$$

Example. The function $i_{\{0\}}$ is the neutral element for infimal convolution: $f \mathbin{\square} i_{\{0\}} = f$ for all functions f.

Example. If $g\colon G \to \mathbf{R}$ is additive, then $f \mathbin{\square} g = g + C$ for some constant $C = C_{f,g} \in [-\infty, +\infty]$. Actually $C_{f,g} = (f \mathbin{\square} g)(0)$.

Example. Define $g_k(x) = k\|x\|$ on a normed space E with norm $\|\cdot\|$, with k a positive constant. Study the infimal convolution $f_k = f \mathbin{\square} g_k$. Assuming that f is bounded we see that f_k is Lipschitz continuous. Moreover $f_k \nearrow f$ as $k \to +\infty$ if f is bounded and lower semicontinuous.

Let us say that a function $f\colon G \to [-\infty, +\infty]$ is *subadditive* if it satisfies the inequality $f(x+y) \leqslant f(x) \dotplus f(y)$, $x, y \in G$.

Proposition 4.3. *A function $f\colon G \to [-\infty, +\infty]$ defined on an abelian group G is subadditive if and only if it satisfies the inequality $f \mathbin{\square} f \geqslant f$. If in addition we assume that $f(0) \leqslant 0$, then this is equivalent to $f \mathbin{\square} f = f$.*

Proof. If f is subadditive we have $f(x-y) \dotplus f(y) \geqslant f(x)$, so taking the infimum over all $y \in G$ gives $f \mathbin{\square} f \geqslant f$. Conversely, we have

$$f(x) \dotplus f(y) \geqslant (f \mathbin{\square} f)(x+y),$$

so $f \mathbin{\square} f \geqslant f$ implies subadditivity. Finally, since the inequality $(f \mathbin{\square} f)(x) \leqslant f(x) \dotplus f(0)$ always holds, we conclude that $f(0) \leqslant 0$ implies $f \mathbin{\square} f \leqslant f$.

Theorem 4.4. *If f and g are subadditive, then so is $f \mathbin{\square} g$.*

Proof. Using the associativity and commutativity of infimal convolution we can write

$$(f \mathbin{\square} g) \mathbin{\square} (f \mathbin{\square} g) = (f \mathbin{\square} f) \mathbin{\square} (g \mathbin{\square} g) \geqslant f \mathbin{\square} g.$$

Subadditive functions are of interest because of their relation to metrics on abelian groups, a fact which we shall discuss now.

Let us call a function $d\colon X \times X \to \mathbf{R}$ a *distance* if it is *positive definite*, i.e., for all $x, y \in X$,

$$d(x,y) \geqslant 0 \text{ with equality precisely when } x = y,$$

and *symmetric*, i.e.,

$$d(x,y) = d(y,x), \qquad x, y \in X.$$

We shall say that a distance is a *metric* if it satisfies in addition the *triangle inequality*, i.e.,

$$d(x,z) \leqslant d(x,y) + d(y,z), \qquad x, y, z \in X.$$

If $X = G$ is an abelian group, *translation-invariant* distances are of interest, i.e., those that satisfy

$$d(x-a, y-a) = d(x,y), \qquad a, x, y \in G.$$

Lemma 4.5. *Any translation-invariant distance on an abelian group G defines a function $f(x) = d(x,0)$ on G which is positive definite,*

$$f(x) \geqslant 0 \textit{ with equality precisely when } x = 0;$$

and symmetric,

$$f(-x) = f(x), \qquad x \in X.$$

Conversely, a function f which is positive definite and symmetric defines a translation-invariant distance $d(x,y) = f(x-y)$.

The proof is easy.

Lemma 4.6. *Let d be a translation-invariant distance on an abelian group G and $f(x) = d(x,0)$. Then d is a metric if and only if f is subadditive,*

$$f(x+y) \leqslant f(x) + f(y), \qquad x, y \in G.$$

Proof. If d is a metric, we can write, using the triangle inequality and the translation invariance,

$$f(x+y) = d(x+y,0) \leqslant d(x+y,y) + d(y,0) = d(x,0) + d(y,0) = f(x) + f(y).$$

Conversely, if f is subadditive,

$$d(x,z) = f(x-z) \leqslant f(x-y) + f(y-z) = d(x,y) + d(y,z).$$

Theorem 4.7 (Kiselman [1996]). *Let $F\colon G \to [0,+\infty]$ be a function on an abelian group G satisfying $F(0) = 0$. Define a sequence of functions $(F_j)_{j=1}^{\infty}$ by putting $F_1 = F$ and $F_j = F_{j-1} \mathbin{\square} F$, $j = 2, 3, \ldots$. Then the sequence (F_j) is decreasing and its limit $\lim F_j = f$ is subadditive. Moreover $\operatorname{dom} f = \mathbf{N} \cdot \operatorname{dom} F$, i.e., f is finite precisely in the semigroup generated by $\operatorname{dom} F$.*

Proof. That the sequence is decreasing is obvious if we take $y = 0$ in the definition of F_{j+1}:

$$F_{j+1}(x) = \inf_y (F_j(x-y) + F(y)) \leqslant F_j(x) + F(0) = F_j(x).$$

Next we shall prove that $f(x+y) \leqslant f(x) + f(y)$. If one of $f(x)$ and $f(y)$ is equal to $+\infty$, there is nothing to prove, so let us assume that $f(x), f(y) < +\infty$ and let us fix a positive number ε. Then there exists numbers j and k such that $F_j(x) \leqslant f(x) + \varepsilon$ and $F_k(y) \leqslant f(y) + \varepsilon$. By associativity $F_{j+k} = F_j \mathbin{\square} F_k$, so we get

$$f(x+y) \leqslant F_{j+k}(x+y) \leqslant F_j(x) + F_k(y) \leqslant f(x) + f(y) + 2\varepsilon.$$

Since ε is arbitrary, the inequality $f(x+y) \leqslant f(x)+f(y)$ follows. The last statement follows from (4.4).

In image analysis it is customary to define distances between adjacent points and then extend the definition to arbitrary pairs of points by going on a path, assigning to each path the sum of the distances between the adjacent points, and finally taking the infimum over all paths. In the translation-invariant case, this amounts to assigning values to a function F at finitely many points, and then define the distance by the function $f = \lim F_j$ of Theorem 4.7. (Of course some extra conditions are needed to ensure that the limit is symmetric and positive definite.) Indeed the paths consists of segments $[0, x^1]$, $[x^1, x^1 + x^2]$,..., $[x + \cdots + x^{k-1}, x]$ and we evaluate the sum $F(x^1) + F(x^2) + \cdots + F(x^k)$ for all possible choices of $x^1, ..., x^k$ with sum x.

Examples of such functions are the following. We always define $F(0) = 0$ and let $F(x) < +\infty$ for x in a finite set P only. The following distances have been studied, assuming P and F to be invariant under permutation and reflection of the coordinates. If we take $P = \{x \in \mathbf{Z}^2;\ \sum |x_j| \leqslant 1\}$ and $F(1,0) = 1$ we get the *city-block distance*, also called l^1. If we let $P = \{x \in \mathbf{Z}^2;\ |x_j| \leqslant 1\}$ and $F(1,0) = F(1,1) = 1$ we get the *chess-board metric*, or l^∞ distance. Other choices are $F(1,0) = a$, $F(1,1) = b$ with $(a,b) = (1, \sqrt{2}), (2,3), (3,4)$. We can also increase the size of P and define $G(1,0) = 5, G(1,1) = 7, G(2,1) = 11$. For references to the work mentioned here see Kiselman [1996].

Proposition 4.8. *Let $f\colon E \to [-\infty, +\infty]$ be a function defined on a vector space E and define f_s by $f_s(x) = sf(x/s)$, $x \in E$, $s \in \mathbf{R} \smallsetminus \{0\}$. Then f is convex if and only if $f_s \mathbin{\square} f_t \geqslant f_{s+t}$ for all $s, t > 0$, and if and only if $f_s \mathbin{\square} f_t = f_{s+t}$ for all $s, t > 0$.*

Proof. We note that the inequality $f_s \mathbin{\square} f_t \leqslant f_{s+t}$ always holds, so the two last properties are indeed equivalent, and the convex functions are those that satisfy the functional equation $f_s \mathbin{\square} f_t = f_{s+t}$; in other words the mapping $s \mapsto f_s$ is a homomorphism of semigroups.

If f is convex, we can write

$$f_s(y) \dotplus f_t(x-y) = sf(y/s) \dotplus tf((x-y)/t) \geqslant (s+t)f\Big(\frac{y}{s+t} + \frac{x-y}{s+t}\Big) = f_{s+t}(x).$$

If we now vary y we obtain $f_s \mathbin{\square} f_t \geqslant f_{s+t}$. Conversely, suppose that this inequality holds. Then we obtain, writing $x_t = (1-t)x_0 + tx_1$, that

$$f(x_t) = f_1(x_t) \leqslant (f_{1-t} \mathbin{\square} f_t)(x_t) \leqslant f_{1-t}(y) \dotplus f_t(x_t - y)$$

for every y. We now choose $y=(1-t)x_0$ and get

$$f(x_t) \leqslant f_{1-t}((1-t)x_0) \dot{+} f_t(x_t-(1-t)x_0) = (1-t)f(x_0) \dot{+} tf(x_1),$$

which means that f is convex.

Theorem 4.9. *If f and g are convex, then so is $f \mathbin{\square} g$.*

Proof. It is easy to verify that $(f \mathbin{\square} g)_s = f_s \mathbin{\square} g_s$. We now perform a calculation like that in the proof of Theorem 4.4:

$$\begin{aligned}(f \mathbin{\square} g)_s \mathbin{\square} (f \mathbin{\square} g)_t = (f_s \mathbin{\square} g_s) \mathbin{\square} (f_t \mathbin{\square} g_t) &= (f_s \mathbin{\square} f_t) \mathbin{\square} (g_s \mathbin{\square} g_t)\\ &\geqslant f_{s+t} \mathbin{\square} g_{s+t} = (f \mathbin{\square} g)_{s+t}.\end{aligned}$$

Thus $f \mathbin{\square} g$ satisfies the criterion of Proposition 4.8 and so is convex.

For a positively homogeneous function convexity is equivalent to subadditivity. This observation will yield a nice formula for the supporting function of the intersection of two closed convex sets, see formula (5.9). Here we note the following easy result.

Proposition 4.10. *Let f, g be two positively homogeneous convex functions. Then $f \mathbin{\square} g$ is the convex hull of their minimum:*

$$\operatorname{cvx}(\min(f,g)) = f \mathbin{\square} g. \tag{4.5}$$

Proof. We always have $f \mathbin{\square} g \leqslant f$, $f \mathbin{\square} g \leqslant g$ since $f(0), g(0) \leqslant 0$. If h is convex and positively homogeneous and $h \leqslant f, g$, then $h = h \mathbin{\square} h \leqslant f \mathbin{\square} g$. Thus $f \mathbin{\square} g$ is the largest positively homogeneous convex minorant of $\min(f,g)$. However, it is easy to see that it is also the largest convex minorant of $\min(f,g)$, whence (4.5).

5. Convex duality: the Fenchel transformation

The affine functions $x \mapsto \xi(x)+c$, where ξ is a linear form and c a real constant, are the simplest convex functions. It is natural to ask whether all convex functions can be somehow represented in terms of these simple functions. The question is analogous to the problem of representing an arbitrary function in Fourier analysis in terms of the simplest functions, the pure oscillations. The Fenchel transformation, which we shall introduce now, plays a role in convexity theory analogous to that of the Fourier transformation in Fourier analysis.

More precisely we ask whether, given a function f on a vector space E, there exists a subset A of $E^* \times \mathbf{R}$ such that

$$f(x) = \sup_{(\xi,c)\in A} \bigl(\xi(x)+c\bigr), \qquad x \in E. \tag{5.1}$$

Here E^* denotes the algebraic dual of E, i.e., the vector space of all linear forms on E. We first note that if this is at all possible, then $c \leqslant f(x)-\xi(x)$ for all $x \in E$ and all $(\xi,c) \in A$, so that $c \leqslant \inf_{x\in E}(f(x)-\xi(x))$. For reasons which will be apparent in a moment, it is convenient to consider instead $-c$; we must have $-c \geqslant \sup_{x\in E}(\xi(x) - f(x))$ for all $(\xi, c) \in A$. We define

$$\widetilde{f}(\xi) = \sup_{x\in E} \big(\xi(x) - f(x)\big), \qquad \xi \in E^*. \tag{5.2}$$

The function $\widetilde{f}$ is called the *Fenchel*[3] *transform* of f. Other names are the *Legendre*[4] *transform* of f and the function *conjugate* to f. Since the constant c in (5.1) must satisfy $c \leqslant -\widetilde{f}(\xi)$, and since $f(x) \geqslant \xi(x) - \widetilde{f}(\xi)$ for all $x \in E$ and all $\xi \in E^*$, we can conclude that (5.1) implies

$$f(x) = \sup_{\xi\in E^*} \big(\xi(x) - \widetilde{f}(\xi)\big), \qquad x \in E.$$

In other words, the supremum in (5.1) does not change if we add points outside A and replace c everywhere by $-\widetilde{f}(\xi)$.

Now the right-hand side of this formula looks like (5.2), so it is natural to apply the transformation a second time. It is convenient here to consider an arbitrary vector subspace F of E^*, and to introduce topologies on E and F as follows. There is a weakest topology on E such that all elements of F are continuous; this is denoted by $\sigma(E, F)$. There is similarly a weakest topology $\sigma(F, E)$ on F such that all evaluation mappings $F \ni \xi \mapsto \xi(x)$, $x \in E$, are continuous. We may for instance choose $F = E^*$, or $F = E'$, the topological dual of E under a given topology, i.e., the space of all continuous linear forms on E. Actually E^* is the topological dual of E equipped with the topology $\sigma(E, E^*)$. Thus, when we speak about the topological dual in the sequel, the case of the algebraic dual is always included as a special case. If E is finite-dimensional and we equip it with the separated vector space topology, then $E' = E^*$. If E is a normed space of infinite dimension, we always have $E' \neq E^*$. It is not necessary that E and F be in duality; we may even choose $F = \{0\}$.

The Fenchel transform of a function g on F is of course a function on the algebraic dual F^* of F:

$$\widetilde{g}(X) = \sup_{\xi\in F} \big(X(\xi) - g(\xi)\big), \qquad X \in F^*. \tag{5.3}$$

Given any element x of E we define an element X of F^* by the formula $X(\xi) = \xi(x)$, $\xi \in F$. Using this idea we may define for any function g on F,

$$\widetilde{g}(x) = \sup_{\xi\in F} \big(\xi(x) - g(\xi)\big), \qquad x \in E. \tag{5.4}$$

We also note that $f \,\square\, \xi$ is an affine function and that in fact $(f \,\square\, \xi)(x) = \xi(x) - \widetilde{f}(\xi)$ for all $x \in E$ and all $\xi \in E^*$. The function $f \,\square\, \xi$ is a minorant of f and in fact the largest affine minorant of f which has linear part equal to ξ. So the supremum of all affine minorants of f with linear part in F is

$$\sup_{\xi\in F}(\xi(x) - \widetilde{f}(\xi)) = \sup_{\xi\in F}(f \,\square\, \xi)(x) = \widetilde{\widetilde{f}}(x), \qquad x \in E.$$

[3] Werner Fenchel, 1905–1988.

[4] Adrien Marie Legendre, 1752–1833.

The question whether (5.1) holds for $A = F \times \mathbf{R}$ can therefore be formulated very succintly: is it true that $\widetilde{\widetilde{f}} = f$? However, it might still be of interest to find a smaller A for which (5.1) holds.

From the definition of $\widetilde{f}$ we immediately obtain the inequality

$$\xi(x) \leqslant f(x) \dotplus \widetilde{f}(\xi), \qquad x \in E, \quad \xi \in E^*, \tag{5.5}$$

called *Fenchel's inequality.* It can be stated equivalently as

$$\xi(x) - \widetilde{f}(\xi) \leqslant f(x), \qquad x \in E, \quad \xi \in E^*.$$

If f is the indicator function of a set A, $f = i_A$, then $\widetilde{f} = H_A$, the supporting function of A. So the question about closed elements for the Galois correspondence studied in the example after Proposition 2.3 will be answered in the more general framework of Fenchel transforms.

We summarize the properties of the Fenchel transformation that we have found so far.

Proposition 5.1. *The Fenchel transformations defined by (5.2) and (5.4) for functions on a vector space E and a subspace F of its algebraic dual form a Galois correspondence, the order of the functions being that of inclusion of their epigraphs. Thus, in terms of the usual order between functions, $f \leqslant g$ implies $\widetilde{f} \geqslant \widetilde{g}$, the second transform satisfies $\widetilde{\widetilde{f}} \leqslant f$, and the third transform is equal to the first, $\big(\widetilde{\widetilde{f}}\big)^{\sim} = \widetilde{f}$. All Fenchel transforms are convex, lower semicontinuous with respect to the topology $\sigma(E, F)$ or $\sigma(F, E)$, and take the value $-\infty$ only when they are identically $-\infty$.*

Proof. Only the last statement does not follow from general properties of Galois correspondences; cf. Proposition 2.1. A supremum of a family of convex functions is convex, in particular so is $\widetilde{f}$. Also the supremum of a family of lower semicontinuous functions is lower semicontinuous. The last property is obvious: if a Fenchel transform $\widetilde{f}$ assumes the value $-\infty$ for a particular ξ, then f must be equal to $+\infty$ identically, and so $\widetilde{f}$ is equal to $-\infty$ identically.

For examples of Fenchel transforms, see for instance Kiselman [1991a or b].

Proposition 5.2. *For any function $f : E \to [-\infty, +\infty]$ on a finite-dimensional vector space the following three conditions are equivalent:*

1. *f is lower semicontinuous, i.e., $\liminf_{y \to x} f(y) = f(x)$ for all $x \in E$;*
2. *$\operatorname{epi} f$ is closed in $E \times \mathbf{R}$;*
3. *For every real number a the sublevel set $\{x \in E;\ f(x) \leqslant a\}$ is closed in E.*

We leave the proof as an exercise.

Theorem 5.3. *Let a function $f : E \to [-\infty, +\infty]$ on a finite-dimensional vector space E be given. Then the following properties are equivalent:*

(A) *f is a Fenchel transform;*

(B) *f is equal to the supremum of all its affine minorants;*

(C) *f is equal to the supremum of some family of affine functions;*

(D) *f is convex, lower semicontinuous, and takes the value $-\infty$ only if it is identically equal to $-\infty$.*

Proof. From Proposition 5.1 and the discussion preceding it is clear that (A), (B), and (C) are all equivalent, and that they imply (D). We need to prove that (D) implies (B), say.

Let us first note that (B) certainly holds if f is either $+\infty$ or $-\infty$. We may therefore suppose that $\operatorname{epi} f$ is nonempty and not equal to the whole space.

We shall prove, assuming that (D) holds, that for any point x_0 the supremum of all affine minorants of f is equal to $f(x_0)$; equivalently, that for any point (x_0, t_0) not in the epigraph of f there is an affine function which takes a value greater than t_0 at x_0. Since $(x_0, t_0) \notin \operatorname{epi} f$ there is a supporting half-space containing $\operatorname{epi} f$ and not containing (x_0, t_0). Such a half-space in $E \times \mathbf{R}$ is defined by an inequality

$$\xi(x) + bt \geqslant c$$

for some $\xi \in E'$ and some real numbers b, c. Since the half-space is supporting, we know that there is some point $(x_1, t_1) \in \operatorname{epi} f$ which satisfies $\xi(x_1) + bt_1 = c$. If $b < 0$, the half-space is the epigraph of an affine function, and since its value at x_0 is larger than t_0, we are done. If $b > 0$, the half-space is the hypograph of an affine function, and it can contain $\operatorname{epi} f$ only if the latter is empty, i.e., $f = +\infty$ identically, a case we already considered. Thus only the case $b = 0$, that of a vertical half-space, remains to be considered. A vertical hyperplane is not the graph of an affine function, and we need to prove that these vertical hyperplanes, although they can occur, do not influence the intersection of all supporting half-spaces.

Thus we have a point (x_0, t_0) not belonging to $\operatorname{epi} f$ and a vertical half-space $\{(x, t) \in E \times \mathbf{R};\ \xi(x) \geqslant c\}$ which contains $\operatorname{epi} f$ and is such that the closest point (x_1, t_1) in $\operatorname{epi} f$ lies in the boundary of the half-space. Then this point must have the same t-coordinate as (x_0, t_0), so $t_1 = t_0$. Since (x_1, t_1) belongs to $\operatorname{epi} f$, the value t_2 of f at x_1 must satisfy $-\infty < t_2 = f(x_1) \leqslant t_1 = t_0$. It is now clear that the closest point in $\operatorname{epi} f$ to the point (x_0, t_2) is (x_1, t_2) and that the hyperplane obtained from our construction is the same, $\xi(x) \geqslant c = \xi(x_1)$. But for points (x_0, t_3) with $t_3 < t_2$ the supporting hyperplane cannot be vertical. Even more interesting is the fact that the supporting hyperplane has a large slope when t_3 is close to t_2. Consider the largest open ball which does not meet $\operatorname{epi} f$ and has its center at (x_0, t_3). Its closure contains a single point of $\operatorname{epi} f$; denote that point by (x_2, t_4) and let R be the radius of the ball. We must have $t_4 > t_3$. The tangent plane to the ball at (x_2, t_4) intersects the line $x = x_0$ at a point $(x_0, t_3 + T)$, where $T = R^2/(t_4 - t_3)$. Since (x_1, t_3) does not belong to $\operatorname{epi} f$ we must have $R > \|x_1 - x_0\|$ and $t_4 > t_3$. On the other hand

$$R = \sqrt{\|x_2 - x_0\|^2 + (t_4 - t_3)^2} \leqslant \sqrt{\|x_1 - x_0\|^2 + (t_2 - t_3)^2},$$

since (x_2, t_4) is the point in $\operatorname{epi} f$ closest to (x_0, t_3) and (x_1, t_2) is a point in $\operatorname{epi} f$. Now $\|x_2 - x_0\| > \|x_1 - x_0\|$, so

$$(t_4 - t_3)^2 \leqslant \|x_1 - x_0\|^2 + (t_2 - t_3)^2 - \|x_2 - x_0\|^2 < (t_2 - t_3)^2.$$

Thus $0 < t_4 - t_3 < t_2 - t_3$. The value at x_0 of the affine function defined by the tangent plane is $t_3 + T$ and tends to plus infinity as $t_3 \nearrow t_2$ since

$$T = \frac{R^2}{t_4 - t_3} > \frac{\|x_1 - x_0\|^2}{t_2 - t_3};$$

in particular $t_3 + T$ is larger than t_0 for some t_3 close to t_2. This proves that the intersection of all supporting half-spaces is not affected if we remove the vertical half-spaces and completes the proof of the theorem.

We now know that the closed elements for the Galois correspondence defined in Proposition 5.1 consists of the functions satisfying condition (D) of Theorem 5.3. An indicator function i_A is closed if and only if the set A is closed and convex. Under the Fenchel transformation these functions are in bijective correspondence with the supporting functions of closed convex sets.

The supporting function mapping $A \mapsto H_A$ embeds the semigroup of all nonempty compact convex subsets of a vector space E into the group of real-valued functions on E'. Thus for instance the equation $A + X = B$, which for nonempty compact convex sets is equivalent to $H_A + H_X = H_B$, can sometimes be solved by a set X, viz. when $H_X = H_B - H_A$ is convex. For applications of the supporting function in image analysis, see Ghosh & Kumar [1998].

Next we shall study the relation between infimal convolution and the Fenchel transformation. The first result is very easy.

Proposition 5.4. *For all functions* $f, g\colon \mathbf{R}^n \to [-\infty, +\infty]$ *we have*

$$(5.6) \qquad (f \mathbin{\square} g)^\sim = \widetilde{f} \mathbin{\underset{\cdot}{+}} \widetilde{g}.$$

In particular, if we take $f = i_X$, $g = i_Y$ *with arbitrary sets* X *and* Y,

$$(5.7) \qquad H_{X+Y} = (i_X \mathbin{\square} i_Y)^\sim = H_X \mathbin{\underset{\cdot}{+}} H_Y.$$

Proof. An easy calculation thanks to the rule (3.3).

We note that we have lower addition in (5.6). We know that $\widetilde{f} \mathbin{\dot{+}} \widetilde{g}$ is convex. However, it turns out that $\widetilde{f} \mathbin{\underset{\cdot}{+}} \widetilde{g} = \widetilde{f} \mathbin{\dot{+}} \widetilde{g}$ except when $\widetilde{f} \mathbin{\underset{\cdot}{+}} \widetilde{g}$ is the constant $-\infty$. Thus also $\widetilde{f} \mathbin{\underset{\cdot}{+}} \widetilde{g}$ is convex. In (5.7) we can write $H_X + H_Y$ without risk of misunderstanding if X and Y are nonempty.

Corollary 5.5. *If* $\widetilde{\widetilde{f}} = f$ *and* $\widetilde{\widetilde{g}} = g$, *then*

$$(5.8) \qquad (f \mathbin{\underset{\cdot}{+}} g)^\sim = (\widetilde{f} \mathbin{\square} \widetilde{g})^{\approx}.$$

In particular, taking $f = i_X$, $g = i_Y$ *with* X *and* Y *closed and convex we have*

$$(5.9) \qquad H_{X \cap Y} = (i_X + i_Y)^\sim = (H_X \mathbin{\square} H_Y)^{\approx} = (\min(H_X, H_Y))^{\approx},$$

and, taking $f = H_X$, $g = H_Y$ *with* X *and* Y *closed and convex,*

$$(5.10) \qquad (H_X \mathbin{\underset{\cdot}{+}} H_Y)^\sim = (i_X \mathbin{\square} i_Y)^{\approx} = (i_{X+Y})^{\approx} = i_{\overline{X+Y}}.$$

Proof. The proof consists of a straightforward application of the proposition, except for the last equality in (5.10), which follows from Theorem 5.3.

More generally, we can obtain the supporting function of an intersection $X = \bigcap_{i\in I} X_i$ of closed convex sets X_i, $i \in I$, as

$$H_X = \Big[\operatorname{cvx}\Big(\inf_{i\in I} H_{X_i}\Big)\Big]^{\widetilde{\ }\widetilde{\ }} = \Big(\inf_{i\in I} H_{X_i}\Big)^{\widetilde{\ }\widetilde{\ }}. \tag{5.11}$$

Theorem 5.6. *Suppose that $\widetilde{\widetilde{f}} = f$, $\widetilde{\widetilde{g}} = g$ and that $\widetilde{f} \,\square\, \widetilde{g}$ is lower semicontinuous and either nowhere minus infinity or else identically minus infinity. Then*

$$(f \mathbin{\underset{\cdot}{+}} g)^{\widetilde{\ }} = \widetilde{f} \,\square\, \widetilde{g}. \tag{5.12}$$

In particular we can take the value at the origin of both sides and obtain

$$-\inf_x (f(x) \mathbin{\underset{\cdot}{+}} g(x)) = \inf_\xi (\widetilde{f}(\xi) \mathbin{\dot{+}} \widetilde{g}(-\xi)). \tag{5.13}$$

Proof. For the proof we only need to combine Corollary 5.5 with Theorem 5.3.

If (5.12) holds, then $\widetilde{f} \,\square\, \widetilde{g}$ is of course lower semicontinuous, so in this respect the result cannot be improved, but it is unpleasant to have lower semicontinuity as an assumption to be verified. Formula (5.12) can be obtained from (5.13) by translation.

We remark that the assumption on $\widetilde{f} \,\square\, \widetilde{g}$ is satisfied if the function is real-valued everywhere, for such functions are automatically continuous as shown by Theorem 3.13. However, in applications it is important to allow the value $+\infty$.

There are important cases when (5.12) does not hold.

Example 5.7. Let $\widetilde{f} = i_X$ and $\widetilde{g} = i_Y$ (then X and Y are automatically closed and convex). Then $\widetilde{f} \,\square\, \widetilde{g} = i_{X+Y}$ and Theorem 5.3 shows that $\big(\widetilde{f} \,\square\, \widetilde{g}\big)^{\widetilde{\ }\widetilde{\ }} = i_{\overline{X+Y}}$, the indicator function of the closure of the convex set $X + Y$. However, $X + Y$ is not necessarily closed. A simple example is

$$X = \{x \in R^2;\, x_2 > 0,\ x_1x_2 \geqslant 1\}, \qquad Y = \{x \in \mathbf{R}^2;\, x_2 = 0\},$$

$$X + Y = \{x \in \mathbf{R}^2;\, x_2 > 0\}, \qquad \overline{X + Y} = \{x \in \mathbf{R}^2;\, x_2 \geqslant 0\}.$$

Then $\widetilde{f} \,\square\, \widetilde{g} \neq \big(\widetilde{f} \,\square\, \widetilde{g}\big)^{\widetilde{\ }\widetilde{\ }}$. The formula (5.12) does not hold.

Example 5.8. Define two functions $f, g\colon \mathbf{R}^2 \to [0, +\infty]$ by

$$f(x) = \begin{cases} 0, & x_1 \leqslant -1, \\ +\infty, & \text{otherwise}; \end{cases} \qquad g(x) = \begin{cases} 0, & x_1 \geqslant 1, \\ +\infty, & \text{otherwise}. \end{cases}$$

Thus $f = i_X$, $g = i_Y$ where X and Y are disjoint closed half-spaces. Then $f \mathbin{\underset{\cdot}{+}} g = +\infty$ identically and $\widetilde{f}(\xi) = H_X(\xi) = -\xi_1$ when $\xi_2 = 0$, $\xi_1 \geqslant 0$ and $+\infty$ otherwise, whereas $\widetilde{g}(\xi) = H_Y(\xi) = \xi_1$ when $\xi_2 = 0$, $\xi_1 \leqslant 0$ and $+\infty$ otherwise. The convex function $\widetilde{f} \,\square\, \widetilde{g}$ takes the value $-\infty$ when $\xi_2 = 0$ and $+\infty$ otherwise. Therefore

$(f \underset{\cdot}{+} g)^{\sim} = \left(\widetilde{f} \,\square\, \widetilde{g}\right)^{\approx} = -\infty$ identically while $\left(\widetilde{f} \,\square\, \widetilde{g}\right)(\xi) = +\infty$ when $\xi_2 \neq 0$. Thus (5.12) does not hold.

The next example is similar to the one we just considered, but in a sense worse, since now $(\widetilde{f} \,\square\, \widetilde{g})(0) = 0$.

Example 5.9. Let

$$X = \{x \in \mathbf{R}^2;\ x_1 > 0,\ x_1 x_2 \geqslant 1\}, \qquad Y = \{x \in \mathbf{R}^2;\ x_1 \leqslant 0\},$$

and consider$f = i_X$, $g = i_Y$. Then $X \cap Y = \emptyset$, so $H_{X\cap Y} = -\infty$ identically. However,

$$\widetilde{f}(\xi) = H_X(\xi) = \begin{cases} -2\sqrt{\xi_1\xi_2}, & \xi_1 \leqslant 0,\ \xi_2 \leqslant 0, \\ +\infty, & \text{otherwise,} \end{cases}$$

and

$$\widetilde{g}(\xi) = H_Y(\xi) = \begin{cases} 0, & \xi_1 \geqslant 0,\ \xi_2 = 0, \\ +\infty, & \text{otherwise,} \end{cases}$$

so that, by Proposition 4.10,

$$(\widetilde{f} \,\square\, \widetilde{g})(\xi) = (H_X \,\square\, H_Y)(\xi) = \operatorname{cvx}(\min(H_X, H_Y))(\xi) = \begin{cases} -\infty, & \xi_2 < 0, \\ 0, & \xi_2 = 0, \\ +\infty, & \xi_2 > 0. \end{cases}$$

In particular we note that $H_{X\cap Y}(0) = -\infty$ while $\operatorname{cvx}(\min(H_X, H_Y))(0) = 0$.

Thus the double tilde in (5.8), (5.9) or (5.10) cannot be omitted in general. The formulas (5.12) and (5.13) are important in optimization, but they are quite subtle—in contrast to (5.6).

Theorem 5.10. *If I is a finite or infinite index set and the A_i are compact convex subsets of a finite-dimensional vector space, then (5.11) can be simplified: the supporting function of the intersection $A = \bigcap A_i$ is*

$$H_A = \operatorname{cvx}\left(\inf_i H_{A_i}\right). \tag{5.14}$$

In particular A is nonempty if and only if

$$\sum H_{A_i}(\xi^i) \geqslant 0 \textit{ for all vectors } \xi^i,\ i \in I, \textit{ such that } \xi^i \neq 0 \\ \textit{for only finitely many indices } i \in I \textit{ and } \sum \xi^i = 0. \tag{5.15}$$

Proof. Let us denote by g the right-hand side of (5.14). Clearly $H_A \leqslant g$. On the other hand, $\widetilde{\widetilde{g}}$ is the supporting function of some set, say $\widetilde{\widetilde{g}} = H_Y$, and this set must be contained in every A_i, thus $Y \subset A$ and $\widetilde{\widetilde{g}} = H_Y \leqslant H_A$. So $\widetilde{\widetilde{g}} \leqslant H_A \leqslant g$; it remains to be proved that $\widetilde{\widetilde{g}} = g$.

If I is empty, both sides of (5.14) equal $+\infty$ identically; if one of the A_i is empty, both sides equal $-\infty$. So let us assume that $I \neq \emptyset$ and that $A_i \neq \emptyset$ for all $i \in I$. Then the H_{A_i} are real-valued and the right-hand side g of (5.14) never takes the value $+\infty$, so $(\operatorname{dom} g)^\circ$ is the whole space. By Theorem 3.13 g is either the constant $-\infty$ or else a continuous real-valued function. So it satisfies $\tilde{\tilde{g}} = g$ in all cases. This proves (5.14). The last statement follows if we use the expression (3.7) for the convex hull of $\inf_i H_{A_i}$ and note that $H_A(0) \geqslant 0$ if and only if $A \neq \emptyset$. This completes the proof.

As shown by Example 5.8, (5.14) does not necessarily hold if the A_i are closed half-spaces and I is finite. Moreover, Example 5.9 shows that (5.15) does not imply that the intersection is nonempty if the A_i are finitely many closed convex subsets. In spite of this, (5.15) does imply that the intersection is nonempty if the A_i are closed half-spaces, finite in number:

Theorem 5.11. *Let A_i, $i \in I$, be finitely many closed half-spaces in a finite-dimensional vector space E. Then their intersection $A = \bigcap A_i$ is nonempty if and only if (5.15) holds; more explicitly, if we assume that the half-spaces are defined by*

$$A_i = \{x \in E;\ \eta^i(x) \leqslant \alpha_i\}, \qquad i \in I, \tag{5.16}$$

for some nonzero linear forms $\eta^i \in E'$ and some real numbers α_i, then (5.15) takes the form

$$\sum_i \lambda\alpha_i \geqslant 0 \textit{ for all } \lambda_i \geqslant 0 \textit{ with } \sum_i \lambda_i\eta^i = 0. \tag{5.17}$$

Proof. The set

$$M = \{(\eta^i, \alpha_i);\ i \in I\} \subset E' \times \mathbf{R}$$

is finite and its convex hull is

$$\operatorname{cvx} M = \Big\{\sum \lambda_i(\eta^i, \alpha_i);\ \lambda_i > 0,\ \sum \lambda_i = 1\Big\}.$$

We define M^+ as the set of all points (ξ, τ') with $\tau' \geqslant \tau$ for some point (ξ, τ) in M. Then $\operatorname{cvx}(M^+) = (\operatorname{cvx} M)^+$. Condition (5.15) means that a point of the form $(0, \tau) \in \operatorname{cvx} M$ must satisfy $\tau \geqslant 0$. Therefore either $0 \notin \operatorname{cvx} M$ or $0 \in \partial(\operatorname{cvx} M)$; we also have $0 \notin (\operatorname{cvx} M)^+$ or $0 \in \partial\big((\operatorname{cvx} M)^+\big)$. There exists a half-space

$$D = \{(\xi, \tau) \in E' \times \mathbf{R};\ \xi(x) + \tau t \leqslant 0\}$$

which contains M^+. Because M^+ contains points with large τ, t must be negative or zero—unless M is empty, but then the conclusion is true anyway. If $t < 0$, the inequality defining D can be written $\xi(-x/t) \leqslant \tau$, and the fact that D contains M can be expressed as $\eta^i(-x/t) \leqslant \alpha_i$, which means that $-x/t \in A$; we are done.

In case $t = 0$ we have a vertical hyperplane in $E' \times \mathbf{R}$. We cannot use the technique in the proof of Theorem 5.3 to tilt the hyperplane, for we are not allowed to lower it, i.e., change the value at the origin of the linear function defining it. But on the other hand, $(\operatorname{cvx} M)^+$ is a polyhedron, which will enable us to use another method: we can tilt the hyperplane and still let it pass through the origin.

So assume that $t = 0$, meaning that we have a vertical half-space

$$D = \{(\xi, \tau);\ \xi(x) \leqslant 0\}$$

containing M. We must then have $x \neq 0$. Consider the set

$$J = \{i \in I;\ \eta^i(x) = 0\}.$$

If J is empty, we have $\eta^i(x) < 0$ for all i and we see that $\eta^i(sx) \leqslant \alpha_i$, thus $sx \in A$, if only s is large enough. Otherwise we consider the problem with points (η^i, α_i), $i \in J$, and the subspace

$$F = \{\xi \in E';\ \xi(x) = 0\}.$$

We may assume that the result is already proved in the space F, which is of smaller dimension. (For spaces of dimension 0 the result is certainly true.) So there exists a point y such that $\eta^i \cdot y \leqslant \alpha_i$ for all $i \in J$. We now have $sx + y \in A$ for s large: if $i \notin J$, then $\eta^i(x) < 0$ and $\eta^i(sx) + \eta^i \cdot y \leqslant \alpha_i$ for $s \gg 0$; if on the other hand $i \in J$, then $\eta^i(sx + y) = \eta^i(y) \leqslant \alpha_i$ by hypothesis. Thus A is nonempty.

Finally, to see that (5.15) takes the form (5.17), it is enough to remark that the supporting functions are

$$H_{A_i}(\xi) = \begin{cases} \lambda\alpha_i \text{ if } \xi = \lambda\eta^i \text{ for some } \lambda \geqslant 0, \text{ and} \\ +\infty \text{ otherwise.} \end{cases}$$

Example 5.12. The conclusion of Theorem 5.11 is not necessarily true if we admit infinite intersections. Let $E = \mathbf{R}^2$ and define infinitely many half-spaces as in (5.16), taking $\eta^i = (1, i) \in \mathbf{R}^2$, $i \in \mathbf{Z}$, and α_i as arbitrary real numbers. Then (5.17) is satisfied regardless of the choice of the α_i. But $A \neq \emptyset$ if and only if

$$\exists C_1\ \exists C_2\ \forall i \in \mathbf{Z} \quad \alpha_i \geqslant -C_1 - iC_2. \tag{5.18}$$

Thus for instance the choice $\alpha_i = \gamma|i|$ yields a nonempty intersection if and only if $\gamma \geqslant 0$. If we take a look at the proof of Theorem 5.11 in this situation, we see that the set M is contained in a vertical half-space $\{(\xi, \tau);\ \xi_1 \geqslant 0\}$. This half-space can be tilted to a non-vertical half-space $\{(\xi, \tau);\ \tau \geqslant -C_1\xi_1 - C_2\xi_2\}$ containing M if and only if $\alpha_i \geqslant -C_1 - iC_2$ for some constants C_1, C_2. (Half-spaces of the form $\{(\xi, \tau);\ \tau \geqslant -C\xi_1 - C_2\xi_2 - \varepsilon\}$ with $\varepsilon > 0$ are not allowed here.) Thus the method of proof we have used works if and only if (5.18) holds.

Maybe we can sum up our experience concerning Theorems 5.10 and 5.11 and the related counterexamples as follows. The calculus of infimal convolution and the Fenchel transformation is highly successful and also quite easy when we consider compact convex sets. When the sets are unbounded, certain difficulties appear—but they have to be confronted! Then again polyhedra with finitely many faces are quite well-behaved even if they happen to be unbounded—but the proofs are quite different!

6. The Monge–Kantorovich problem

We shall now discuss an application of convex duality, the Monge–Kantorovich[5] problem. It is about moving masses of earth around in the most economical way. But the masses could also be images. Kantorovich's theory has applications in economics and image analysis.

Given two probability measures as finite sums of Dirac measures

$$\mu = \sum A_j \delta_{a_j} \text{ and } \nu = \sum B_k \delta_{b_k}, \tag{6.1}$$

the *Kantorovich cost functional* is

$$C(\mu, \nu) = \inf_M \sum m_{jk} c(a_j, b_k), \tag{6.2}$$

where $c(a, b)$ denotes the cost in Euro to move one ton of earth from place a to place b, or the cost in öre to move one pixel from a to b on the screen, and where the infimum is taken over all matrices $M = (m_{jk})$ such that $\sum_k m_{jk} = A_j$ (moving the mass A_j out from a_j to the various b_k) and $\sum_j m_{jk} = B_k$ (moving the mass B_k to b_k from all possible a_j).

On a screen we may have $512 \times 1024 = 2^{19}$ pixels, so M is a matrix with 2^{38} entries. Therefore the problem is unwieldy. Thomas Kaijser [1998] has studied it and devised algorithms to calculate the functional.

Kantorovich let the measures be arbitrary probability measures and defined the cost (or work) as an integral.

We can approximate the A_j and B_k by rational numbers, and then we can even assume that they are all equal to $1/m$ for some m, for the locations a_j and b_k can be repeated at will. This is the situation we shall consider here; it can be viewed as an approximation to the general problem.

So let us consider a metric space X with metric d and linear combinations of Dirac measures

$$\mu = \frac{1}{m} \sum_{j=1}^{m} \delta_{a_j}, \qquad \nu = \frac{1}{m} \sum_{j=1}^{m} \delta_{b_j}, \tag{6.3}$$

where $m \in \mathbf{N} \smallsetminus \{0\}$ and a_j and b_j are points in X, $j = 1, ..., m$. The points could represent pixels in an image, and the number of indices j such that a_j is equal to a particular point represents the brightness of the image at that point.

Let us define a distance d_1 for such measures by putting

$$d_1(\mu, \nu) = \inf_\sigma \frac{1}{m} \sum_{j=1}^{m} d(a_j, b_{\sigma(j)}), \tag{6.4}$$

where the infimum is taken over all permutations σ of $\{1, 2, ..., m\}$. In this case M is equal to $1/m$ times a permutation matrix, and the cost $c(a, b)$ is a metric $d(a, b)$; the

[5] Gaspar Monge, 1746–1818; Leonid Vital′evič Kantorovič, 1912–1986.

triangle inequality will be needed. (It is probably of interest to consider other cost functionals, but then they can hardly be equal to the dual distance defined below.) Thus d_1 measures the work needed to move μ to ν: we choose to move each a_j to some b_k and then take the most economical permutation. This is the *Kantorovich distance* between μ and ν. With $m = 2^{19}$ pixels, the infimum in (6.4) is taken over $(2^{19})!$ permutations—it is indeed unwieldy.

The *dual distance* d_2 between two measures is defined by

$$d_2(\mu,\nu) = \sup_{f\in \mathrm{Lip}} |\mu(f) - \nu(f)|, \tag{6.5}$$

where Lip denotes the set of all Lipschitz functions with Lipschitz constant 1; i.e., functions $f\colon X \to \mathbf{R}$ such that

$$|f(x) - f(y)| \leqslant d(x,y), \qquad x, y \in X.$$

Theorem 6.1 (Kantorovich [1942]). *For any two measures as in (6.3) we have $d_1(\mu,\nu) = d_2(\mu,\nu)$.*

This is Kantorovich's classical theorem restricted to this special case, which, however, can be easily extended to the general case by approximating arbitrary probability measures by sums of Dirac measures. We shall present a new proof here using convex duality, more precisely Theorem 5.11.

Proof. It is easy to see that $d_2 \leqslant d_1$. To prove the inequality in the other direction we shall construct a function $f = \min_j f_j \in \mathrm{Lip}$, where

$$f_j(x) = c_j + d(x, b_j), \qquad x \in X,\ j = 1, ..., m,$$

for some skilfully chosen real numbers c_j. We want that, after some permutation of the b_j, $f(x) = f_j(x)$ for $x = a_j, b_j$. If we succeed in this construction, we will have $f(a_j) - f(b_j) = d(a_j, b_j)$ for all j and we shall obtain

$$d_2(\mu,\nu) \geqslant |\mu(f) - \nu(f)| = \frac{1}{m}\sum (f(a_j) - f(b_j)) = \frac{1}{m}\sum d(a_j, b_j) \geqslant d_1(\mu,\nu),$$

which will finish the proof.

Thus the question is to find levels c_j so that $f(a_j) = f_j(a_j)$ and $f(b_j) = f_j(b_j)$. Here the first equality holds if and only if $f_k(a_j) \geqslant f_j(a_j)$ for all k, i.e., $c_k + d(a_j, b_k) \geqslant c_j + d(a_j, b_j)$. The second holds if and only if $f_k(b_j) \geqslant f_j(b_j)$ for all k, i.e., $c_k + d(b_j, b_k) \geqslant c_j$. We note that the second condition follows from the first, for if the first is satisfied, then $c_k - c_j \geqslant d(a_j, b_j) - d(a_j, b_k) \geqslant -d(b_j, b_k)$ in view of the triangle inequality. So let us forget about the second condition. We introduce the numbers $\theta_{jk} = d(a_j, b_k)$. Our task is to find numbers c_j such that $c_k - c_j \geqslant \theta_{jj} - \theta_{jk}$ for all j, k. Such numbers can be found if and only if the θ_{jk} satisfy the condition (6.7) below.

Proposition 6.2. *Given real numbers θ_{jk}, $j, k = 1, ..., m$, there exists numbers c_j, $j = 1, ..., m$, such that*

$$c_k - c_j \geqslant \theta_{jj} - \theta_{jk}, \qquad j, k = 1, ..., m, \tag{6.6}$$

if and only if

$$\sum_1^m \theta_{jj} \leqslant \sum_1^m \theta_{j,\sigma(j)} \text{ \textit{for all permutations} } \sigma \text{ \textit{of} } \{1, ..., m\}. \tag{6.7}$$

Proof. It is clear that (6.7) follows from (6.6). For the converse we note that we have a purely geometric problem in $\mathbf{R}^m$ with m^2 closed half-spaces

$$A_{jk} = \{c \in \mathbf{R}^m;\ c_j - c_k \leqslant \theta_{jk} - \theta_{jj}\}; \tag{6.8}$$

we wish to show that their intersection A is nonempty. We shall use Theorem 5.11 and the criterion (5.17); thus A is nonempty if and only if (5.17) holds. We now apply this criterion to the situation we have in the proposition. We have m^2 half-spaces (6.8), where, in the notation of (5.16), $\eta^{jk} = e^{(j)} - e^{(k)}$ and $\alpha_{jk} = \theta_{jk} - \theta_{jj}$. We conclude that the intersection A is nonempty if and only if

$$\sum_{jk} \lambda_{jk}(\theta_{jk} - \theta_{jj}) \geqslant 0 \text{ for all } \lambda_{jk} \geqslant 0 \text{ such that } \sum_k \lambda_{sk} = \sum_j \lambda_{js} \text{ for all } s. \tag{6.9}$$

Since we can add diagonal matrices freely to (λ_{jk}) without changing either the assumption or the conclusion in (6.9), it is enough that (6.9) be satisfied for bistochastic matrices (λ_{jk}). It is even enough to require it for a special type of bistochastic matrices, viz. the permutation matrices, for the convex hull of all permutation matrices is precisely the set of bistochastic matrices. This is because the permutation matrices are the extremal points of the set of all bistochastic matrices, and a well-known theorem states that a compact convex set is the closed convex hull of its extremal points. But when (λ_{jk}) is a permutation matrix, (6.9) reduces to (6.7). This concludes the proof the proposition.

Proof of Theorem 6.1, cont'd. Is condition (6.7) satisfied in the situation of Theorem 6.1? If we put $\theta_{jk} = d(a_j, b_k)$, then (6.7) becomes

$$\sum d(a_j, b_j) \leqslant \sum d(a_j, b_{\sigma(j)}).$$

There exists a permutation π such that

$$\sum d(a_j, b_{\pi(j)}) = \inf_\sigma \sum d(a_j, b_{\sigma(j)}).$$

We can now renumber the points b_j so that π becomes the identity. Then (6.7) holds, and we have proved the theorem.

7. The Brunn–Minkowski inequality

In my lectures on the Brunn–Minkowski inequality and the Prékopa–Leindler inequality I followed Ball [1997] closely. Therefore I do not include anything from these lectures here.

8. Non-convex sets

Can we extend the techniques used in convexity theory to non-convex sets? Here I shall indicate very briefly how it is possible to use the supporting function of a convex set to describe faithfully also non-convex closed sets.

Proposition 8.1. *Consider the mapping* $p\colon x \mapsto \big(x, \frac12\|x\|^2\big)$ *from* $\mathbf{R}^n$ *into* $\mathbf{R}^{n+1}$, *where* $\|\cdot\|$ *is a Euclidean norm. The inverse image of the convex hull of the image of any set in* $\mathbf{R}^n$ *is equal to the set itself:*

$$p^{-1}\big(\operatorname{cvx} p(A)\big) = A, \qquad A \subset \mathbf{R}^n. \tag{8.1}$$

For closed subsets A *of* $\mathbf{R}^n$ *we can take the closed convex hull:*

$$p^{-1}\big(\overline{\operatorname{cvx} p(A)}\big) = A, \qquad A \textit{ closed in } \mathbf{R}^n. \tag{8.2}$$

Proof. The mapping p is an embedding of $\mathbf{R}^n$ into $\mathbf{R}^{n+1}$, preserves the C^∞ structure of $\mathbf{R}^n$, and realizes $\mathbf{R}^n$ as the paraboloid $P = p(\mathbf{R}^n) = \{(x,t) \in \mathbf{R}^{n+1};\ t = \frac12\|x\|^2\}$. Also P is the graph of the mapping $\varphi\colon x \mapsto \frac12\|x\|^2$, which has the property that $\operatorname{grad}\varphi = \mathrm{id}_{\mathbf{R}^n}$.

Clearly A is contained in $p^{-1}(\operatorname{cvx} p(A))$. To prove the other inclusion, observe that the tangent plane to P at a point $(a,s) \in P$ is $\{(x,t);\ t - s = a\cdot(x-a)\}$. The closed half-space $\{(x,t);\ t - s \geqslant a\cdot(x-a)\}$ contains the paraboloid. If $a \notin A$, then the corresponding open half-space $D = \{(x,t);\ t - s > x\cdot(x-a)\}$ contains $p(A)$ and therefore also its convex hull. Now D does not contain (a,s), so $p^{-1}(D)$ does not contain a. This proves (8.1).

If A is closed and $a \notin A$, then there is even a closed half-space of the form $D_\varepsilon = \{(x,t);\ t - s \geqslant a\cdot(x-a) + \varepsilon\}$ for some positive ε which contains $p(A)$, hence also its closed convex hull. More precisely, if $(x,t) \in P \cap D_\varepsilon$, then $\|x-a\| \geqslant \sqrt{2\varepsilon}$. Hence $a \notin p^{-1}\big(\overline{\operatorname{cvx} p(A)}\big)$. This proves (8.2).

The supporting function of the paraboloid P is given by

$$H_P(\xi,\tau) = \begin{cases} -\frac{1}{2\tau}\|\xi\|^2, & \tau < 0,\\ +\infty, & \tau > 0 \text{ or } \tau = 0,\ \xi \neq 0,\\ 0, & (\xi,\tau) = (0,0). \end{cases}$$

The closed convex hull of the image of a set A is the intersection of all closed half-spaces containing $p(A)$, thus

$$\overline{\operatorname{cvx} p(A)} = \bigcap_{\xi,\tau}\{(x,t) \in \mathbf{R}^{n+1};\ \xi\cdot x + \tau t \leqslant H_{p(A)}(\xi,\tau)\}.$$

Hence the inverse image of $\overline{\operatorname{cvx} p(A)}$ is

$$p^{-1}\big(\overline{\operatorname{cvx} p(A)}\big) = \bigcap_{\xi,\tau}\{x \in \mathbf{R}^n;\ \xi\cdot x + \tfrac12\tau\|x\|^2 \leqslant H_{p(A)}(\xi,\tau)\}. \tag{8.3}$$

Therefore any closed set A can be recovered from the supporting function of $p(A)$.

We can simplify (8.3): it is not necessary to take the intersection over all $(\xi,\tau) \in \mathbf{R}^n \times \mathbf{R}$. With the notation

$$E_\tau = \bigcap_{\xi}\{(x,t) \in \mathbf{R}^n \times \mathbf{R};\ \xi\cdot x + \tau t \leqslant H_{p(A)}(\xi,\tau)\},$$

we can write $\overline{\operatorname{cvx} p(A)} = \bigcap_\tau E_\tau = E_1 \cap E_0 \cap E_{-1}$, for $E_\tau = E_1$ if τ is positive, $E_\tau = E_{-1}$ if τ is negative. I claim that $E_{-1} \cap P = E_1 \cap E_0 \cap E_{-1} \cap P$. Indeed, if (x,t) belongs to $E_{-1} \cap P$, then x belongs to $\overline{A}$, as shown by the proof of Proposition 8.1, for there we used the half-spaces D_ε which have $\tau = -1$. And if $x \in \overline{A}$, then clearly $p(x)$ belongs to all E_τ. Therefore it is enough to take $\tau = -1$ in the intersection in (8.3).

Finally we note that $\overline{p(A)} = p\big(\overline{A}\big)$ and that $\overline{\operatorname{cvx} p(A)} = \overline{\operatorname{cvx}\big(\overline{p(A)}\big)}$ (cf. Proposition 2.3), so that $\overline{A} = p^{-1}(\overline{\operatorname{cvx}(p(A)})$; cf. (8.2). We may now write

$$\overline{A} = p^{-1}\big(\overline{\operatorname{cvx} p(A)}\big) = \bigcap_\xi \{x \in \mathbf{R}^n;\ \xi \cdot x - \tfrac{1}{2}\|x\|^2 \leqslant H_{p(A)}(\xi, -1)\}.$$

We also note that $H_{p(A)}(a,-1) \leqslant H_P(a,-1)$ with equality precisely when $a \in \overline{A}$. We sum up the discussion as follows.

Proposition 8.2. *Define for any subset A of $\mathbf{R}^n$ a function*

$$\Gamma_A(\xi) = \sup_{x \in A} \big(\xi \cdot x - \tfrac{1}{2}\|x\|^2\big) = H_{p(A)}(\xi, -1), \qquad \xi \in \mathbf{R}^n.$$

Then $\Gamma_A(a) \leqslant \frac{1}{2}\|a\|^2$ everywhere, with equality if and only if a belongs to the closure of A.

This idea can be used to recover the support, not just the convex hull of the support, from the Fourier transform of a distribution with compact support; see Kiselman [1981].

9. Notions of topology

9.1. Mappings

Let $f\colon X \to Y$ be a mapping from a set X into a set Y, and denote by $P(X)$ the *power set* of X, i.e., the set of all subsets of X. We associate with f a mapping $f^*\colon P(Y) \to P(X)$ and a mapping $f_*\colon P(X) \to P(Y)$ defined as follows.

(9.1.1)
$$f^*(B) = \{x \in X;\ f(x) \in B\} = f^{-1}(B),\ \ f_*(A) = \{f(x) \in Y;\ x \in A\} = f(A).$$

Thus $f^*(B) = f^{-1}(B)$ is the *preimage* (*inverse image*) of $B \subset Y$ and $f_*(A) = f(A)$ is the *direct image* (or just *image*) of $A \subset X$. It is however sometimes convenient to have a special notation for $f_*\colon P(X) \to P(Y)$, so that it is not confused with $f\colon X \to Y$; similarly f^* is not the pointwise inverse of f.

We note that

$$f^* \circ f_* \geqslant \mathrm{id}_{P(X)} \text{ and } f_* \circ f^* \leqslant \mathrm{id}_{P(Y)}. \tag{9.1.2}$$

Thus $f^*(f_*(A)) \supset A$; equality holds for all A if and only if f is injective, and $f_*(f^*(B)) = B \cap \operatorname{im} f \subset B$; equality holds for all B if and only if f is surjective.

We also note that f^* is a homomorphism of Boolean algebras: it satisfies

$$f^*(B_1 \cup B_2) = f^*(B_1) \cup f^*(B_2), \qquad f^*(B_1 \cap B_2) = f^*(B_1) \cap f^*(B_2), \tag{9.1.3}$$

$$(9.1.4) \qquad f^*(B_1 \smallsetminus B_2) = f^*(B_1) \smallsetminus f^*(B_2), \text{ in particular } f^*\big(\complement B\big) = \complement f^*(B).$$

More generally, (9.1.3) can be generalized to infinite unions and intersections. The homomorphism f^* is an endomorphism if and only if f is surjective, and an epimorphism if and only if f is injective.

The mapping f_* is not so well-behaved: it always satisfies

$$f_*(A_1 \cup A_2) = f_*(A_1) \cup f_*(A_2),$$

but only $f_*(A_1 \cap A_2) \subset f_*(A_1) \cap f_*(A_2)$, and there is in general no inclusion relation between $f_*\big(\complement A\big)$ and $\complement f_*(A)$.

9.2. Definition of topologies

A *topology* on a set X is a collection $\mathcal{U}(X)$ of subsets of X, thus an element of $P(P(X))$, which is stable under arbitrary unions and finite intersections. The elements of $\mathcal{U}(X)$ are called *open sets*; thus any union of open sets is open and any finite intersection of open sets is open. In particular, the union and the intersection of the empty family is open, so $\varnothing$ and X are always open subsets of X.

However, a topology can be given in several different ways. We define a set as *closed* if its complement is open. Then the family $\mathcal{F}(X)$ of all closed sets is stable under arbitrary intersections and finite unions. We may also impose these conditions as axioms, and define a set to be open if its complement is closed. A topology can be equivalently defined using open or closed sets.

Another notion is that of neighborhood. If a topology $\mathcal{U}(X)$ is given, we say that a set V is a *neighborhood* of a point x if there exists an open set U such that $x \in U \subset V$. The families $\mathcal{V}(x)$, $x \in X$, of neighborhoods of points in X satisfy the following conditions:

$$(9.2.1) \qquad \text{If } V \in \mathcal{V}(x), \text{ then } x \in V;$$

$$(9.2.2) \qquad \text{If } V \in \mathcal{V}(x) \text{ and } W \supset V, \text{ then } W \in \mathcal{V}(x);$$

$$(9.2.3) \qquad \text{If } V_1, V_2 \in \mathcal{V}(x), \text{ then } V_1 \cap V_2 \in \mathcal{V}(x);$$

$$(9.2.4) \quad \text{If } V \in \mathcal{V}(x), \text{ then there exists } W \in \mathcal{V}(x) \text{ such that } V \in \mathcal{V}(y) \text{ for all } y \in W.$$

These properties are easy to verify if the topology is given and the neighborhoods are defined as above. On the other hand, if we have a collection $\mathcal{V}(x)$ for every $x \in X$ satisfying the axioms (9.2.1)–(9.2.4) and define a set U to be open if it belongs to $\mathcal{V}(x)$ for every $x \in U$, then we get a topology for which the neighborhoods are the given ones.

We can also define a topology using closure operators. If a topology is given, then we can define a closure operator by taking $\overline{A}$ as the intersection of all closed sets containing A. Then this closure operator satisfies $\overline{\varnothing} = \varnothing$ and $\overline{A \cup B} = \overline{A} \cup \overline{B}$. Conversely, if a closure operator is defined satisfying these conditions we can define a set to be closed if $\overline{A} = A$; we then get a topology, a topology for which the topological closure operator is the given one.

Finally, the interior A° of a set A is the largest open set contained in the set. It is related to the closure by the formula

$$A^\circ = \complement\left(\overline{\complement A}\right).$$

The operation $A \mapsto A^\circ$ is shrinking, increasing (order preserving), and idempotent:

$$A^\circ \subset A;$$

$$A_1 \subset A_2 \text{ implies } A_1^\circ \subset A_2^\circ; \text{ and}$$

$$(A^\circ)^\circ = A^\circ.$$

This means that it is a closure operator if we reverse the order: $A \leqslant B$ shall mean $A \supset B$. In addition to being a closure operator, it satisfies $X^\circ = X$ and $(A \cap B)^\circ = A^\circ \cap B^\circ$. Conversely, we may take these properties as axioms and define a set to be open if it is in the image of the operator. Then we get a topology and the operation of taking the interior of a set for this topology is equal to the original operator.

Summing up, we have five equivalent ways to define a topology: using open sets, closed sets, neighborhoods, taking the topological closure, and taking the interior.

If we have two topologies $\mathcal{U}_1(X)$ and $\mathcal{U}_2(X)$ on the same set X we say that the first is *weaker* or *coarser* than the second, and that the second is *finer* or *stronger* than the first, if $\mathcal{U}_1(X) \subset \mathcal{U}_2(X)$. Expressed in terms of closure operators, this means that $c_2 \leqslant c_1$ if c_j denotes the closure operator associated with $\mathcal{U}_j(X)$, $j = 1, 2$. The weakest topology is the *chaotic topology* $\{\varnothing, X\}$ and the strongest is the *discrete topology* $P(X)$. The closure of a nonempty set in the chaotic topology is always the whole space, wheras the closure of a set in the discrete topology is the set itself.

A two-point space can have four topologies: in addition to the two just mentioned, they are $\{\varnothing, \{x\}, \{x, y\}\}$ and $\{\varnothing, \{y\}, \{x, y\}\}$. The two latter are called *Sierpiński*[6] *topologies.* How many topologies are there on a three-point space?

9.3. Transport of topologies

If $f\colon X \to Y$ is a mapping from a set X into a topological space Y we can transport the topology on Y to X by defining a subset of X to be open if and only if it is of the form $f^*(U)$ for some open subset U of Y. Because of (9.1.3) and the corresponding formula for infinite unions it is clear that the family of all sets

$$(f^*)_*(\mathcal{U}(Y)) = \{f^*(U);\, U \in \mathcal{U}(Y)\}$$

is a topology. Here we have used the notation introduced in (9.1.1) at the next higher level: $f^*\colon P(Y) \to P(X)$, $(f^*)_*\colon P(P(Y)) \to P(P(X))$. For brevity we shall denote $(f^*)_*(\mathcal{U}(Y))$ by $f^{\leftarrow}(\mathcal{U}(Y))$, the *pull-back* of $\mathcal{U}(Y)$.

If $d\colon Y \to Y$ is a closure operator in Y, then $d^{\leftarrow} = f^* \circ d \circ f_*$ is a closure operator in X, and if d satisfies the topological axioms $d(\varnothing) = \varnothing$ and $d(B_1 \cup B_2) = d(B_1) \cup d(B_2)$, then $d^{\leftarrow}$ does the same. Thus we can transport topological closure operators from Y

[6] Wacław Sierpiński, 1882–1969.

to X. One can verify that the transported open sets correspond to the transported closure operator.

A particularly common case is when X is a subset of Y and f is the inclusion mapping. Then we say that the topology $f^{\leftarrow}(\mathcal{U}(Y))$ defined on X is the *induced topology.* We see that U is open in X if and only if $U = V \cap Y$ for some open set in Y; we also see that the closure operator $d^{\leftarrow}$ in X is defined as $d^{\leftarrow}(A) = d(A) \cap X$.

If X is a topological space and $f\colon X \to Y$ a mapping of X into a set Y, we can of course consider the family $\{f_*(A);\ A \in \mathcal{U}(X)\}$. However, since f_* is not so well-behaved, it is usually not a topology on Y. Instead we use again f^* and declare a subset B of Y to be open if $f^*(B)$ is open in X. And we can verify that this is indeed a topology on Y; we shall denote it by

$$f_{\rightarrow}(\mathcal{U}(X)) = (f^*)^*(\mathcal{U}(X)) = \{B \in P(Y);\ f^*(B) \in \mathcal{U}(X)\},$$

the *push-forward* of $\mathcal{U}(X)$.

A common instance of this definition is when Y is a quotient set of X, i.e., when we have an equivalence relation $\sim$ in X and let $Y = X/{\sim}$ be the set of all equivalence classes in X with respect to the relation. The mapping f associates to each element in X its equivalence class in Y. Then a subset B of Y is open in Y with respect to the topology we have pushed forward from X if and only if the union of all equivalence classes in B is open in X. The topology obtained in this way on $X/{\sim}$ is called the *quotient topology.*

If $f\colon X \to Y$ is injective, and if we have a topology on X, push it forward to Y and then pull it back to X, the new topology agrees with the original one: $f^{\leftarrow}(f_{\rightarrow}(\mathcal{U}(X))) = \mathcal{U}(X)$. Similarly, if f is surjective and we start with a topology on Y, pull it back to X and then push it forward to Y, we obtain the original topology; $f_{\rightarrow}(f^{\leftarrow}(\mathcal{U}(Y))) = \mathcal{U}(Y)$. (This works so well because we did not use f_* but f^* in the definition.)

However, if we have a closure operator c in X, we cannot define a closure operator in Y by something like $c_{\rightarrow} = f_* \circ c \circ f^*$. In general $c_{\rightarrow}$ will not be expanding, nor idempotent. (Construct examples!) How shall we define the closure operator connected with the topology $f_{\rightarrow}(\mathcal{U}(X))$ on Y?

9.4. Continuous mappings

Let $f\colon X \to Y$ be a mapping of a topological space X into a topological space Y and x a point in X. We say that f is *continuous at x* if $f^*(V)$ is a neighborhood of x for every neighborhood V of $f(x)$. It is called *continuous* if it is continuous at every point in X. We now translate this well-known notion into the language of open sets, closed sets, and closure operators. Then we can prove that f is continuous if and only if $f^*(U) \in \mathcal{U}(X)$ for every $U \in \mathcal{U}(Y)$; in other words if and only if the topology $f^{\leftarrow}(\mathcal{U}(Y))$ is weaker than the topology $\mathcal{U}(X)$.

9.5. Connectedness

The family of all open and closed sets of a topological space X (sometimes called the *clopen* sets) forms a Boolean algebra. This algebra must contain the two sets $\varnothing, X$, for they are always both open and closed. (If X is empty there is only one such set, of course.) A topological space is said to be *connected* if it is nonempty and the only sets which are both open and closed are the empty set and the whole space.

A subset of a topological space is called *connected* if it is connected as a topological space with the induced topology.[7] A *connectivity component* of a topological space is a connected subset which is maximal with respect to inclusion.

A connected subset which is both open and closed is a component. It is easy to prove that the closure of a connected subset is connected. Therefore all components are closed. They need not be open.

Proposition 9.5.1. *Let $f\colon X \to Y$ be a continuous mapping of a topological space X into a topological space Y. If X is connected, then so is $f_*(X) = \operatorname{im} f$.*

Proof. Let B be a clopen subset of $\operatorname{im} f$. Then $f^*(B)$ is clopen in X. Hence $f^*(B)$ is either empty or equal to X. Therefore $f_*(f^*(B)) = B \cap \operatorname{im} f = B$ is either empty or equal to $\operatorname{im} f$. This means that $\operatorname{im} f$ is connected.

Corollary 9.5.2. *Let $f\colon X \to Y$ be a mapping of a topological space X into a set Y. Equip Y with the strongest topology such that f is continuous. Suppose that X is connected. Then $\operatorname{im} f$ is connected, and the points in $Y \smallsetminus \operatorname{im} f$ are isolated. In particular, any quotient space of a connected topological space is connected.*

Proof. For any point $y \in Y$ not in the image of f, the inverse image $f^{-1}(y) = f^*(\{y\})$ is empty, thus both open and closed. This means that $Y \smallsetminus \operatorname{im} f$ has the discrete topology and the connectivity components are just the singleton sets.

In particular we shall use Corollary 9.5.2 with $X = \mathbf{R}$ and $Y = \mathbf{Z}$ to define connected topologies on the digital line $\mathbf{Z}$. Let $f\colon \mathbf{R} \to \mathbf{Z}$ be a surjective mapping. Then $\mathbf{Z}$ equipped with the strongest topology such that f is continuous is a connected topological space. Thus we consider $\mathbf{Z}$ as a quotient space of $\mathbf{R}$, not as a subspace. It is not unnatural to restrict attention to increasing mappings $f\colon \mathbf{R} \to \mathbf{Z}$. For every $n \in \mathbf{Z}$ we then have two numbers $a_n < b_n$ such that

$$]a_n, b_n[\subset f^*(n) \subset [a_n, b_n].$$

We can normalize the situation to $a_n = n$, $b_n = n + 1$; this does not change the topology on $\mathbf{Z}$. Then $f(x) = \lfloor x \rfloor$ for all $x \in \mathbf{R} \smallsetminus \mathbf{Z}$, and $f(n) = n$ or $f(n) = n - 1$, $n \in \mathbf{Z}$. The topology is therefore determined if we know for which n we have $f(n) = n - 1$. For every subset A of $\mathbf{Z}$ we get a topology on $\mathbf{Z}$ by declaring that $f(n) = n - 1$ for $a \in A$ and that, for all other real numbers x, we have $f(x) = \lfloor x \rfloor$.

Another normalization is to take $a_n = n - \frac{1}{2}$, $b_n = n + \frac{1}{2}$. This can be explained as follows. It is natural to think of $\mathbf{Z}$ as an approximation of the real line $\mathbf{R}$ and to consider mappings $f\colon \mathbf{R} \to \mathbf{Z}$ expressing this idea. We may define $f(x)$ to be the integer closest to x; this is well-defined unless x is a half-integer. So when $x = n + \frac{1}{2}$ we have a choice for each n: shall we define $f(n+\frac{1}{2}) = n$ or $f(n+\frac{1}{2}) = n+1$? If we choose the first alternative for every n, thus putting $f^*(n) = \left]n - \frac{1}{2}, n + \frac{1}{2}\right]$, the topology defined in Corollary 9.5.2 is called the *right topology* on $\mathbf{Z}$; if we choose the second, we obtain the *left topology* on $\mathbf{Z}$; cf. Bourbaki [1961:I:§1: Exerc. 2]. Another choice is

[7] According to Bourbaki [1961:I:§11:1] the empty space is connected. Here I follow instead the advice of Adrien Douady (personal communication, June 26, 2000). In these notes it will not matter whether the empty set is said to be connected or not.

to always choose an even integer as the best approximant of a half-integer. Then the closed interval $[-\frac{1}{2}, \frac{1}{2}]$ is mapped to 0, so $\{0\}$ is closed, whereas the inverse image of 1 is the open interval $]\frac{1}{2}, \frac{3}{2}[$, so that $\{1\}$ is open. This topology was introduced by E. D. Halimskiĭ (Efim Khalimsky), and we shall call it the *Khalimsky topology*; $\mathbf{Z}$ with this topology is called the *Khalimsky line*. The Khalimsky line is connected, but the complement of any point is disconnected. Among all the topologies defined by increasing surjections $f\colon \mathbf{R} \to \mathbf{Z}$ only two have this property: the one just defined and the one obtained by translating everything by one step.

9.6. Separation axioms and adjacency

The closure of a subset A of a topological space X will be denoted by $\overline{A}$. The intersection of all neighborhoods of a point x will be denoted by $N(x)$. We note that $x \in \overline{\{y\}}$ if and only if $y \in N(x)$. The relation $x \in \overline{\{y\}}$ defines a preorder in X, i.e., a relation satisfying (2.1) and (2.3) but not necessarily (2.2). We shall denote it by $x \preccurlyeq y$; thus $x \preccurlyeq y$ if and only if $x \in \overline{\{y\}}$. It was introduced by Aleksandrov [1937:503]. We shall call it the *specialization preorder*; cf. Kong et al. [1991:905].

A *Kolmogorov space* (Bourbaki [1961:I:§1: Exerc. 2]), also called a T_0-space, is a topological space such that $x \in N(y)$ and $y \in N(x)$ only if $x = y$, thus precisely when the specialization preorder is an order (satisfies (2.2)). Conversely, every ordered set can be made into a T_0-space by defining the smallest neighborhood of a point x to be $N(x) = \{y \in X;\ x \preccurlyeq y\}$. It is quite reasonable to impose this axiom; if x belongs to the closure of $\{y\}$ and vice versa, then x and y are indistinguishable from the point of view of topology. (We should therefore identify them and consider a quotient space.)

The separation axiom T_1 states that $N(x) = \{x\}$. It is too strong to be of interest for the spaces considered here. Also the specialization preorder becomes uninteresting: we have $x \preccurlyeq y$ if and only if $x = y$.

Two points x and y in a topological space Y are said to be *adjacent* if $x \neq y$ and $\{x, y\}$ is connected. We note that $\{x, y\}$ is connected if and only if either $x \in N(y)$ or $y \in N(x)$. We shall say that two points x, z are *second adjacent* if $x \neq z$; x and z are not adjacent; and there exists a third point $y \in Y$ such that x and y are adjacent and y and z are adjacent.

10. Smallest neighborhood spaces

In a topological space the union of any family of open sets is open. It may happen that also the intersection of any family of open sets is open. Equivalently, every point in the space possesses a smallest neighborhood. A space with this property we shall call here a *smallest neighborhood space*. Another suitable name would be a *P. S. Aleksandrov space*, in honor of P. S. Aleksandrov,[8] who introduced them in [1935, 1937]. It is equivalent to require that the union of an arbitrary family of closed sets is closed.

The intersection $N(x)$ of all neighborhoods of a point x is open for all x if and only if the space is a smallest neighborhood space.

Aleksandrov [1935, 1937] introduced the term *espace discret, diskreter Raum* (*discrete space*) for a topological space such that the intersection of any family of

[8] Pavel Sergeevič Aleksandrov, 1896–1982; not to be confused with Aleksandr Danilovič Aleksandrov, b. 1912.

open sets is open. The intersection of all closed sets containing a set M he called its *Hülle* (*hull*), and denoted it by $\overline{M}$ or AM. The intersection of all open sets containing a set M he called its *Stern* (*star*) and denoted it by OM. He noted that the star of a set is a closure operation, and therefore defines a topology, which he called *réciproque* [1935] or *dual* [1937]. The closed set of a smallest neighborhood space satisfies the axioms of the open sets of a topology, so there is a complete symmetry between the two topologies in such a space.

Alexandrov's choice of terms seems fortunate, but nowadays it is not possible to use the term *discrete space* in Aleksandrov's sense, since the discrete topology in modern usage refers only to the topology where every set is open, the strongest of all topologies. This is why I propose to call a discrete space in Aleksandrov's sense a *smallest neighborhood space* or a *P. S. Aleksandrov space*.

The closed points, i.e., the points x such that $\overline{\{x\}} = \{x\}$, Aleksandrov called *Eckpunkte* (*vertices*), and the open points, i.e., the points x such that the singleton $\{x\}$ is open, he called *Grundpunkte* (*base points*).

We can define a topology on the digital line $\mathbf{Z}$ by declaring all odd points to be open, thus $N(2k+1) = \{2k+1\}$, and all even points to have a smallest neighborhood $N(2k) = \{2k-1, 2k, 2k+1\}$. It follows that the even points are closed, for the complement of an even point $2k$ is the union of all $N(x)$ with $x \neq 2k$, thus an open set. This is the Khalimsky topology already defined in Section 9.5. Thus in the Khalimsky topology the even points are *Eckpunkte* and the odd points are *Grundpunkte* in Aleksandrov's terminology. In the specialization order, the base points are higher than the vertices...

A *Khalimsky interval* is an interval $[a, b] \cap \mathbf{Z}$ equipped with the topology induced by the Khalimsky topology on $\mathbf{Z}$. A *Khalimsky circle* is a quotient space $\mathbf{Z}_m = \mathbf{Z}/m\mathbf{Z}$ of the Khalimsky line for some even integer $m \geqslant 4$. (If m is odd, the quotient space receives the chaotic topology, which is not interesting.)

The *Khalimsky plane* is the Cartesian product of two Khalimsky lines, and more generally, *Khalimsky n-space* is the Cartesian product of n copies of $\mathbf{Z}$. Equivalently, we can define Khalimsky n-space on $\mathbf{Z}^n$ by declaring $\{x \in \mathbf{Z}^n;\ \|x - c\|_\infty \leqslant 1\}$ to be open for any point $c \in (2\mathbf{Z})^n$ and then taking all intersections of such sets as open sets, then all unions of such intersections.

There are, however, other topologies in $\mathbf{Z}^2$ which are of interest: we may declare $\{x \in \mathbf{Z}^2;\ \|x - c\|_1 \leqslant 1\}$ to be open for any c such that $\sum c_j \in 2\mathbf{Z}$ as well as all intersections of such sets.[9] The Khalimsky topology and the topology just defined are not comparable: none is stronger than the other. However, they are related, for if we turn the Khalimsky plane 45° and delete all points which are not open or closed, we obtain the new topology; see the proof of Theorem 11.3.1.

11. Digital Jordan curve theorems

The classical Jordan curve theorem says that the complement of a Jordan curve in the Euclidean plane $\mathbf{R}^2$ consists of exactly two connectivity components. Efim

[9] I found this topology in response to a question asked by Timur Sadykov on February 1, 2000. However, I found out later that it was defined already by Wyse [1970]; see Section 11.3.

Khalimsky's digital Jordan curve theorem states the same thing for the digital plane $\mathbf{Z}^2$. Of course we must use a suitable definition of the concept of digital Jordan curve, as well as a suitable topology on $\mathbf{Z}^2$. In this case $\mathbf{Z}^2$ is given the Cartesian product topology of two copies of the digital line $\mathbf{Z}$ equipped with the Khalimsky topology.

A proof of Khalimsky's theorem was published in 1990 by Khalimsky, Kopperman and Meyer [1990]. They refer to earlier proofs by Khalimsky (E. D. Halimskiĭ) [1970, 1977]. We shall present a new, short proof here.

The idea of the proof is simple. For the smallest Jordan curves (having four or eight points) the conclusion of the theorem can be proved by inspection. Given any other Jordan curve J, we construct a Jordan curve J' which has shorter Euclidean length and is such that its complement has as many components as the complement of J. Since the possible Euclidean lengths form a discrete set, this procedure will lead to one of the smallest Jordan curves, for which the theorem is already established. The construction of J' can intuitively be described as follows: attack J where its curvature is maximal and shorten it there; it cannot offer resistance from within.

We then consider a topology on $\mathbf{Z}^2$ which is not a product topology. In contrast to the Khalimsky topology it has the property that every point is either open or closed. We prove that the Jordan curve theorem holds for this topology for a restricted class of Jordan curves.

11.1. Khalimsky Jordan curves

Khalimsky, Kopperman and Meyer [1990: 3.1] used the following definitions of path and arc in the Khalimsky plane. We just extend them here to any topological space. We modify slightly their definition of a Jordan curve [1990: 5.1]. A Jordan curve in the Euclidean plane $\mathbf{R}^2$ is a homeomorphic image of the circle $\mathbf{R}/\mathbf{Z}$, and similarly a Khalimsky Jordan curve is a homeomorphic image of a Khalimsky circle.

Definition 11.1.1. *Let Y be any topological space. A* ***Khalimsky path*** *in Y is a continuous image of a Khalimsky interval. A* ***Khalimsky arc*** *is a homeomorphic image of a Khalimsky interval. A* ***Khalimsky Jordan curve*** *in Y is a homeomorphic image of a Khalimsky circle.*

Sometimes Khalimsky Jordan curves are too narrow. We impose a condition on them to make their interior fatter:

Definition 11.1.2. *Let J be a Khalimsky Jordan curve in a topological space Y. We shall say that J is* ***strict*** *if every point in J is second adjacent to exactly two points in J.*

We note that if $x, z \in J$ are second adjacent, then the intermediary y required by the definition need not belong to J. Thus the concept of strict Jordan curve is not intrinsic.

A three-set $\{x, y, z\}$ such that all three points are adjacent to each other can be a Khalimsky path but never a Khalimsky arc. This follows from the fact that in a Khalimsky interval $[a, b]$, the endpoints a and b are not adjacent unless $b = a + 1$. Let us say that a three-set $\{x, y, z\}$ in a topological space is a *forbidden triangle* if all points are adjacent to each other. The absence of forbidden triangles is therefore a necessary condition for Khalimsky arcs and consequently for Khalimsky Jordan curves, and it is often easy to check.

Different topologies may induce the same adjacency structure. However, when the adjacency structure is that of a Khalimsky circle, the topology of the space must also be that of a Khalimsky circle. More precisely we have the following result.

Theorem 11.1.3. *Given a subset J of a topological space Y, the following conditions are equivalent.*
(A) J is a Khalimsky Jordan curve.
(B) J has at least four points, and for every $a \in J$, $J \smallsetminus \{a\}$ is homeomorphic to a Khalimsky interval.
(C) J is finite, connected, with cardinality at least 4, and each of its elements has exactly two adjacent points.
(D) J has the adjacency structure of a Khalimsky circle, i.e., $J = \{x_1, x_2, ..., x_m\}$ for some even integer $m \geqslant 4$ and for each $j = 1, ..., m$, x_{j-1} and x_{j+1} and no other points are adjacent to x_j. (Here we count indices modulo m.)

Proof. If (A) holds, then for every $a \in J$, $J \smallsetminus \{a\}$ is homeomorphic to a Khalimsky circle minus one point, thus to a Khalimsky interval. Conversely, suppose that (B) holds and consider $J \smallsetminus \{a\}$ and $J \smallsetminus \{b\}$ for two points $a, b \in J$ which are not adjacent. Then we have homeomorphisms of $J \smallsetminus \{a\}$ and $J \smallsetminus \{b\}$ into a Khalimsky circle $\mathbf{Z}_m$. We can modify them by rotating the circle so that the two mappings agree on $J \smallsetminus \{a, b\}$. Then they define a local homeomorphism of J onto $\mathbf{Z}_m$, thus a homeomorphism; we have proved (A).

It is clear that (A) implies (C) and (D).

Suppose that (C) holds. Then call an arbitrary point x_1 and one of its adjacent points x_2 and then go on, always choosing x_{j+1} after x_j so that x_{j+1} is adjacent to x_j but not equal to any of the already chosen $x_1, ..., x_{j-1}$. After a while we must arrive at a situation where there are no new points left, i.e., we arrive at x_m and the two points adjacent to x_m are x_{m-1} and a point which has already been chosen, say x_k. A priori k may be any of $1, 2, ..., m-2$, but in fact the only possibility is $k = 1$—any other choice would mean that x_k had three adjacent points contrary to the assumption. It remains to be seen that m is even. That x_j and x_{j+1} are adjacent means that we have either $x_j \in N(x_{j+1})$ or $x_{j+1} \in N(x_j)$. If $x_j \in N(x_{j+1})$, then we cannot have $x_{j+1} \in N(x_{j+2})$, for that would imply that x_j belonged to $N(x_{j+2})$, so that x_{j+2} would have three adjacent elements, viz. x_j, x_{j+1} and x_{j+3}. So the statement $x_j \in N(x_{j+1})$ holds only for j of a certain parity. Since this is true modulo m, that number must be even. Thus we have proved (D). Conversely, (D) obviously implies (C) since (D) is just a more detailed version of (C).

It remains to be seen that (D) implies (A). First of all it is clear that, assuming (D), $N(x)$ can never have more than three elements—a fourth element would mean that x had at least three adjacent points. So $N(x_j) \subset \{x_{j-1}, x_j, x_{j+1}\}$. Considering the three points x_{j-1}, x_j, x_{j+1}, we note that either $x_{j-1} \in N(x_j)$ or $x_j \in N(x_{j-1})$, and that $x_j \in N(x_{j+1})$ or $x_{j+1} \in N(x_j)$. However, these alternatives cannot be chosen at will, for as we have seen in the previous paragraph $x_{j-1} \in N(x_j)$ implies $x_j \notin N(x_{j+1})$. Consider now the case $x_{j-1} \in N(x_j)$. Then $x_{j+1} \in N(x_j)$, so that $N(x_j) \supset \{x_{j-1}, x_j, x_{j+1}\}$. On the other hand we know already that $N(x_j)$ has at most three elements; we conclude that $N(x_j) = \{x_{j-1}, x_j, x_{j+1}\}$. By the same argument, $N(x_{j+2}) = \{x_{j+1}, x_{j+2}, x_{j+3}\}$. Therefore $N(x_{j+1}) = \{x_{j+1}\}$, and we have proved that Y is a Khalimsky circle where points with indices of the same parity as j have three-neighborhoods and points with indices of the other parity are open. The other possibility, viz. that $x_j \in N(x_{j-1})$, can be reduced to the former by just shifting the indices one step.

It follows from property (C) that two Khalimsky Jordan curves can never be contained in each other. More precisely, if J and K are Khalimsky Jordan curves and $J \subset K$, then $J = K$.

A point on a Khalimsky Jordan curve J consisting of at least six points has at least two second adjacent points; with the order introduced in property (D), x_{j-2} and x_{j+2} are second adjacent to x_j and $x_{j-2} \neq x_{j+2}$ when $m > 4$. Then $x_{j\pm1}$ serve as intermediaries, but there may also exist other intermediaries. When a Jordan curve is not strict and $m > 4$, then some point, say x_j, has at least one second adjacent point in addition to x_{j-2} and x_{j+2}, say x_k. Then an intermediary b such that x_j and b are adjacent and b and x_k are adjacent cannot belong to J.

Suppose now that Y is a metric space with metric d. Since every Khalimsky arc Γ is homeomorphic either to $[0, m-1] \cap \mathbf{Z}$ or to $[1, m] \cap \mathbf{Z}$ for some m, it can be indexed as $\{x_1, ..., x_m\}$, where the indices are uniquely determined except for inversion. We may define its length as

$$\operatorname{length}(\Gamma) = \sum_1^{m-1} d(x_{j+1}, x_j).$$

Similarly, a Khalimsky Jordan curve can be indexed as $\{x_1, ..., x_m\}$, where the indices are uniquely determined up to inversion and circular permutations, and its length can be defined as

$$\operatorname{length}(J) = \sum_1^{m} d(x_{j+1}, x_j),$$

where we count the indices modulo m.

We shall use the following norms in $\mathbf{R}^2$ to measure distances in $\mathbf{Z}^2$:

$$\|x\|_p = \|(x_1, x_2)\|_p = \begin{cases} \big(|x_1|^p + |x_2|^p\big)^{1/p}, & x \in \mathbf{R}^2, \quad 1 \leqslant p < +\infty; \\ \max\big(|x_1|, |x_2|\big), & x \in \mathbf{R}^2, \quad p = \infty. \end{cases}$$

11.2. Khalimsky's digital Jordan curve theorem

The Khalimsky topology of the digital plane is the Cartesian product topology of two copies of the Khalimsky line $\mathbf{Z}$. A point $x = (x_1, x_2)$ in the product $\mathbf{Z}^2 = \mathbf{Z} \times \mathbf{Z}$ is closed if and only if both x_1 and x_2 are closed, thus if and only if both x_1 and x_2 are even; similarly x is open if and only if both coordinates are odd. These points are called *pure*; the other points, which are neither open nor closed, are called *mixed.*

Perhaps the quickest way to describe Khalimsky's topology τ_∞ on $\mathbf{Z}^2$ is this: We first declare the nine-set

(11.2.1) $\quad U_\infty = \{x \in \mathbf{Z}^2;\ \|x\|_\infty \leqslant 1\} = \{(0,0), \pm(1,0), \pm(1,1), \pm(0,1), \pm(-1,1)\}$

to be open, as well as all translates $U_\infty + c$ with $c_1, c_2 \in 2\mathbf{Z}$. Then all intersections of such translates are open, as well as all unions of the sets so obtained. As a consequence, $\{(1,-1),(1,0),(1,1)\}$, the intersection of U_∞ and $U_\infty + (2,0)$, and $\{(1,1)\}$, the intersection of U_∞ and $U_\infty + (2,2)$, are open sets, and $\{(0,0)\}$ is a closed set. The sets $\{(1,0)\}$ and $\{(0,1)\}$ are neither open nor closed.

Theorem 11.2.1. *Given a subset J of $\mathbf{Z}^2$ equipped with the Khalimsky topology, the conditions A, B, C and D of Theorem 11.1.3 are all equivalent to the following.*
(E) $J = \{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$ for some even integer $m \geqslant 4$ and for all j, $x^{(j-1)}$ and $x^{(j+1)}$ and no other points are adjacent to $x^{(j)}$; moreover each path consisting of three consecutive points $\{x^{(j-1)}, x^{(j)}, x^{(j+1)}\}$ turns at $x^{(j)}$ by 45° or 90° or not at all if $x^{(j)}$ is a pure point, and goes straight ahead if $x^{(j)}$ is mixed.

Here we use the informal expression "turn by 45°" etc. with reference to angles in the Euclidean plane of which we consider the Khalimsky plane to be a subset (not a subspace).

Proof. If (D) holds, we see that J cannot turn at a mixed point and cannot turn 135° at a pure point—otherwise we would have a forbidden triangle. So (D) implies (E). Conversely, (E) is just a more precise version of (D), so (E) implies (D).

In this section we shall measure the lengths of Khalimsky Jordan curves using the Euclidean metric, $d(x,y) = \|x-y\|_2$. It is not possible to use $\|\cdot\|_1$ or $\|\cdot\|_\infty$ in the proof of the Jordan curve theorem.

The smallest possible Jordan curve in $\mathbf{Z}^2$ is the four-set

$$J_4 = \{x \in \mathbf{Z}^2;\ \|x-(1,0)\|_1 = 1\} = \{(0,0),(1,-1),(2,0),(1,1)\}.$$

We add all translates of J_4 by a vector $c \in \mathbf{Z}^2$ with $c_1 + c_2$ even and call these the *Jordan curves of type J_4*.

There is also a Jordan curve having eight points,

$$J_8 = \{x \in \mathbf{Z}^2;\ \|x\|_\infty = 1\} = U_\infty \smallsetminus \{(0,0)\}. \tag{11.2.2}$$

This curve and all its translates by a vector $c \in \mathbf{Z}^2$ with $c_1 + c_2$ even we call the *Jordan curves of type J_8*.

Let us agree to call the three-set

$$T = \{(1,1),(0,0),(1,-1)\} \tag{11.2.3}$$

and rotations of T by 90°, 180° and 270°, as well as all translates of these sets by vectors $c \in \mathbf{Z}^2$ with $c_1 + c_2$ even, a *removable triangle*. It turns out that elimination of removable triangles is a convenient way to reduce Jordan curves, as shown by the following lemma.

Lemma 11.2.2. *Let J be a Jordan curve in the Khalimsky plane and assume that J contains the three-set T defined by (11.2.3). Define*

$$J' = (J \smallsetminus \{(0,0)\}) \cup \{(1,0)\}.$$

Then either $J = J_4$ or else J' is a Jordan curve such that $\complement J'$ and $\complement J$ have the same number of components, and length(J') = length$(J) - 2\sqrt{2} + 2$.

Proof. Assume first that $(2,0) \in J$, thus that $J \supset J_4$. Then necessarily $J = J_4$.

Next we suppose that $(2,0) \notin J$. Then J' is a Jordan curve: J' is a set where the new point $(1,0)$ plays exactly the same role topologically as the old point $(0,0)$ in J. Thus J' is also homeomorphic to a Khalimsky circle.

Finally we must check that the number of components in $\complement J'$ is the same as that of $\complement J'$. Indeed, $(1,0)$ and $(2,0)$ belong to the same component of $\complement J$, and $(0,0)$ and $(-1,0)$ belong to the same component of $\complement J'$.

Theorem 11.2.3 (Khalimsky's Jordan curve theorem). *Let us equip the digital plane* $\mathbf{Z}^2$ *with the Khalimsky topology* τ_∞ *(see (11.2.1)). Then for any Khalimsky Jordan curve* J *in* $\mathbf{Z}^2$*, the complement* $\complement J = \mathbf{Z}^2 \smallsetminus J$ *has exactly two connectivity components.*

Proof. The complement of J_4 consists of $A = \{(1,0)\}$ and the set B of all points x with $|x_1 - 1| + |x_2| > 1$. It is obvious that these two sets are connected. Moreover, A is closed and open in $\complement J_4$, so it is a component. Therefore, also B is closed and open in $\complement J_4$ and also a component. The proof for J_8 is similar.

Thus we know that the conclusion of the theorem holds for Jordan curves of types J_4 and J_8.

Next we shall prove that if J is not of the kind already treated, then there exists a Jordan curve J' of strictly smaller Euclidean length such that $\complement J$ and $\complement J'$ have the same number of components. After a finite number of steps we must arrive at a situation where the hypothesis is no longer satisfied, which means that we have a Jordan curve of type J_4 or J_8, for which the complement has two components as we already proved.

The construction of J' is as follows. First we may assume, in view of Lemma 11.2.2, that J contains no removable triangles. Define

$$a_2 = \inf(x_2; x \in J).$$

Thus $x_2 \geqslant a_2$ for all points $x \in J$ with equality for at least one x. Consider a horizontal interval

$$H = \{(x_1, a_2)\} + \{(0,0), (1,0), ..., (p,0)\}$$

which is maximal with respect to inclusion and consists of points in J with ordinate equal to a_2. The maximality implies that the two points $(x_1 - 1, a_2)$ and $(x_1 + p + 1, a_2)$ do not belong to J. Then we see that p must be an even number, but we cannot have $p = 0$, since that would imply that J contained a removable triangle, contrary to the assumption. Thus H contains at least three points. Moreover, at the endpoints of H, J must turn upwards. Indeed, since $(x_1 - 1, a_2)$ does not belong to J, exactly one of the points $(x_1 - 1, a_2 + 1)$, $(x_1, a_2 + 1)$ belongs to J; when we go left from (x_1, a_2), the curve must turn upwards by either 45° or 90°; it cannot turn downwards. Similarly, the curve turns upwards by 45° or 90° when we go right from the last point in H, viz. from $(x_1 + p, a_2)$.

We now consider the set $\mathfrak{I}$ of all maximal horizontal intervals I in J such that J turns upwards at the endpoints of I. The previous argument served just to prove that there exists such an interval. Now there exists an interval $K \in \mathfrak{I}$ of smallest length,

$$K = \{y\} + \{(0,0), (1,0), ..., (q,0)\},$$

containing $q+1$ points for some even number $q \geqslant 2$. We shall assume that K is of smallest length also among all intervals that can be obtained from the intervals in $\mathcal{J}$ by rotating them 90°, 180° or 270°.

To simplify the notation we may assume (after a translation if necessary) that $y=(0,0)$, so that

$$K = \{(0,0),(1,0),...,(q,0)\} = [(0,0),(q,0)] \cap \mathbf{Z}^2.$$

Case 1. J turns upwards by 45° at both ends of K. This means that $(-1,1)$ and $(q+1,1)$ both belong to J. In this case, we define

$$J' = (J \smallsetminus K) \cup (K+(0,1)).$$

This operation shortens the Euclidean length by $2\sqrt{2}-2$ (but it does not shorten the l^∞ length). We note that the interval $K+(0,1)$ is disjoint from J; otherwise some point in K would have three adjacent points. Moreover $K+(0,2)$ must be disjoint from J. Indeed, if $(K+(0,2)) \cap J$ were nonempty, then either J would contain a removable triangle (contrary to our assumption) or there would exist a subinterval K' of $K+(0,2)$ contained in J and such that J turns upwards at its endpoints; thus $K' \in \mathcal{J}$. This subinterval must have fewer than $q+1$ points, since $(0,2)$ and $(q,2)$ cannot belong to J—otherwise there would be a removable triangle in J. Now a shorter interval is impossible, since K is by assumption an interval in $\mathcal{J}$ of shortest length. One checks that J' is a Jordan curve. Indeed, the points of $K+(0,1)$ play the same role topologically in J' as do the points of K in J. The number of components in the complement of J' is the same as for J.

Case 2. J turns upwards by 90° at one end of K. Assume that $(0,1) \in J$, the case $(q,1) \in J$ being symmetric. Then also $(0,2) \in J$. We consider the subcases 2.1 and 2.2.

Case 2.1. $(2,2) \notin J$. We cut off a corner, i.e., we remove $(0,1)$, $(0,0)$, $(1,0)$, and add $(1,1)$. This operation shortens the Euclidean length by $4-2\sqrt{2}$ (but J' has the same l^1-length as J). Since $(1,1)$ and $(2,2)$ belong to the same component of $\complement J$, and $(0,1)$, $(0,0)$, $(1,0)$, and $(-1,0)$ belong to the same component of $\complement J'$, the number of components in the respective complements are the same.

Case 2.2. $(2,2) \in J$. We consider four subcases, 2.2.1.1, 2.2.1.2, 2.2.2.1 and 2.2.2.2.

Case 2.2.1.1. $(2,1) \in J$, $(1,2) \in J$. Then J contains a Jordan curve of type J_8, more precisely $J \supset (1,1)+J_8$. So J must be equal to that curve.

Case 2.2.1.2. $(2,1) \in J$, $(1,2) \notin J$. Remove the five points $(0,1)$, $(0,0)$, $(1,0)$, $(2,0)$, $(2,1)$, and add $(1,2)$. Thus J' is shorter by 4. We can check that J' has all desired properties.

Case 2.2.2.1. $(2,1) \notin J$, $(1,2) \in J$. Turn 90° to reduce to case 2.2.1.2.

Case 2.2.2.2. $(2,1) \notin J$, $(1,2) \notin J$. This case cannot occur since q is smallest possible. To see this, define I' as the set of all points $(2,2),(3,2),...,(q',2) \in J$ with q' as large as possible. If J turns upwards at $(q',2)$, then I' belongs to $\mathcal{J}$ with $q' < q$, which contradicts the definition of K and q. If on the other hand J turns downwards at $(q',2)$, then there exists a vertical interval consisting of three points, which becomes an interval in $\mathcal{J}$ if we turn it 90°, thus again contradicting the definition of $\mathcal{J}$.

11.3. The Jordan curve theorem for another topology

We define a topology τ_1 on $\mathbf{Z}^2$ by first declaring the five-set

$$U_1 = \{x \in \mathbf{Z}^2;\ \|x\|_1 \leqslant 1\} = \{(0,0), \pm(1,0), \pm(0,1)\} \tag{11.3.1}$$

to be open, then all translates $U + c$ with $c \in \mathbf{Z}^2$, $c_1 + c_2 \in 2\mathbf{Z}$ to be open, as well as all intersections of such translates. This implies that $\{(1,0)\}$ is open, and that the origin is closed. In fact, all points $x \in \mathbf{Z}^2$ with $x_1 + x_2 \in 2\mathbf{Z}$ are closed, and all points with $x_1 + x_2 \notin 2\mathbf{Z}$ are open; there are no mixed points. This topology was described by Wyse et al. [1970] and Rosenfeld [1979: 624].

The four-set

$$J'_4 = \{(0,0), (1,0), (1,1), (0,1)\}$$

is a Jordan curve for τ_1. However, it is not strict, for a point in J'_4 has only one second adjacent point. Its complement is connected, so the Jordan curve theorem does not hold. The set J_8 defined by (11.2.2) is a Khalimsky Jordan curve and its complement has exactly two components. Also J_8 is not strict, because the point $(1,0)$ has three second adjacent points, viz. $(0,1)$, $(0,-1)$ and $(-1,0)$.

Another example is the twelve-set

$$J_{12} = \{(0,0), (1,0), (2,0), (2,1), (3,1), (3,2), (3,3), (2,3), (1,3), (1,2), (0,2), (0,1)\}.$$

It is a Jordan curve, not strict, and its complement has three connectivity components, viz. an infinite component and the two singleton sets $\{(1,1)\}$ and $\{(2,2)\}$.

Theorem 11.3.1. *Let $\mathbf{Z}^2$ be equipped with the topology τ_1 just defined (see (11.3.1)). Then the complement of every strict Jordan curve has exactly two components.*

Proof. For the proof we shall use the fact that $\mathbf{Z}^2$ equipped with the topology τ_1 is homeomorphic to the subspace of all pure points in the Khalimsky plane. This fact was used also by Kong, Kopperman and Meyer [1991: 915].

Let X be the digital plane $\mathbf{Z}^2$ with the topology τ_1, and Y the Khalimsky plane ($\mathbf{Z}^2$ with the topology τ_∞). Consider the mapping $\varphi\colon X \to Y$ defined by $\varphi(x) = (x_1 - x_2, x_1 + x_2)$. Its image $\varphi(X)$ is the set of all pure points in Y, and if we equip it with the topology induced by Y it is homeomorphic to X. Moreover, the image of any Khalimsky Jordan curve J in X is a Khalimsky Jordan curve in Y. Therefore $Y \smallsetminus \varphi(J)$ has exactly two components by Theorem 11.2.3. We claim that $\varphi(X) \smallsetminus \varphi(J)$ has exactly two components. It is clear that this set has at least two components, so the problem is to prove that a component A of $Y \smallsetminus \varphi(J)$ gives rise to a connected set $A \cap \varphi(X)$, i.e., that the pure points in A form a connected set.

To this end, assume that $a, a' \in A \cap \varphi(X)$, and consider a Khalimsky arc

$$\{a = a^{(0)}, a^{(1)}, ..., a^{(s)} = a'\}$$

contained in $Y \smallsetminus \varphi(J)$. (Connectedness in Y is the same as arcwise connectedness; cf. Khalimsky et al. [1990: Theorem 3.2].) We shall prove that this arc can be replaced by another consisting only of pure points. So assume that $a^{(j)}$ is a mixed point. Then its predecessor $a^{(j-1)}$ and its successor $a^{(j+1)}$ are both pure points. Without loss of generality, we may assume that $a^{(j-1)} = (0,0)$, $a^{(j)} = (0,1)$, and $a^{(j+1)} = (0,2)$.

We may then replace $a^{(j)}$ by one of the pure points $(-1,1)$, $(1,1)$, because both of them cannot belong to $\varphi(J)$. To see this, suppose that $(1,1), (-1,1) \in \varphi(J)$. Then $(-1,1)$ would be a second adjacent point to $(1,1)$, and this point has, by hypothesis, exactly two second adjacent points in $\varphi(J)$ (considering everything in the space $\varphi(X)$). However, none of them can be equal to $(-1,1)$, for the only possible intermediaries would then be $(0,0)$ and $(0,2)$, none of which belongs to $\varphi(J)$. (In a strict Jordan curve, one of the possible intermediaries to a second adjacent point must belong to the curve.) This contradiction shows that not both of $(1,1)$ and $(-1,1)$ can belong to $\varphi(J)$. Thus we may define $b = (-1,1)$ or $b = (1,1)$ so that $b \notin \varphi(J)$ and observe that

$$\{a^{(0)}, ..., a^{(j-1)}, b, a^{(j+1)}, ..., a^{(s)}\}$$

is a Khalimsky arc with a mixed point replaced by a pure point. After finitely many such replacements we obtain an arc connecting a and a' and consisting only of pure points. This shows that $\varphi(X) \smallsetminus \varphi(J)$ has at most as many components as $Y \smallsetminus \varphi(J)$; therefore exactly two components, and then the same is true of $X \smallsetminus J$.

References

Alexandroff, Paul [Aleksandrov, P. S.] (1896–1982)

1935 Sur les espaces discrets. *C. R. Acad. Sci. Paris* **200**, 1649–1651.

1937 Diskrete Räume. *Mat. Sb.* **2** (44), 501–519.

Ball, Keith

1997 An elementary introduction to modern convex geometry. In: *Flavors of Geometry*, 1–58. Math. Sci. Res. Inst. Publ. 31. Cambridge: Cambridge University Press.

Bourbaki, Nicolas

1961 *Topologie générale*. Éléments de mathématique, première partie, livre III, chapitres 1 & 2. Third edition. Hermann.

Ghosh, Pijush K.; Kumar, K. Vinod

1998 Support function representation of convex bodies, its application in geometric computing, and some related representations. *Computer Vision and Image Understanding* **72**, 379–403.

Halimskiĭ, E. D.

1970 Applications of connected ordered topological spaces in topology. Conference of Math. Departments of Povolsia.

1977 *Uporyadochennnye topologicheskie prostranstva.* Kiev: Naukova Dumka. 92 pp.

Hiriart-Urruty, Jean-Baptiste; Lemaréchal, Claude

1993 *Convex Analysis and Minimization Algorithms I. Fundamentals.* Springer-Verlag, VXII + 417 pp.

Kaijser, Thomas

1998 Computing the Kantorovich distance for images. *Journal of Mathematical Imaging and Vision* **9**, 173–191.

Kantorovich, L.

1942 On the translocation of masses. *Comptes Rendus (Doklady) de l'Académie des Sciences de l'URSS* **37**, 199–201.

Khalimsky, Efim; Kopperman, Ralph; Meyer Paul R.

1990a Computer graphics and connected topologies on finite ordered sets. *Topol. Appl.* **36**, 1–17.

1990b Boundaries in digital planes. *J. Appl. Math. Stoch. Anal.* **3**, 27–55.

Kiselman, Christer O.

1981 How to recognize supports from the growth of functional transforms in real and complex analysis. In: *Functional analysis, Holomorphy, and Approximation Theory* (Ed. S. Machado); Lecture Notes in Mathematics **843**, pp. 366–372. Springer-Verlag.

1991a *Konvexa mängder och funktioner.* Uppsala University, Department of Mathematics, Report No. 1991:LN1. 34 pp.

1991b *Konveksaj aroj kaj funkcioj.* Uppsala University, Department of Mathematics, Report No. 1991:LN2. 35 pp.

1996 Regularity properties of distance transformations in image analysis. *Computer Vision and Image Understanding*, **64**, No. 3, 390–398.

2000 Digital Jordan curve theorems. *Discrete Geometry for Computer Imagery*, 9th International Conference, DGCI 2000, Uppsala, Sweden, December 13–15, 2000. (Eds. Gunilla Borgefors, Ingela Nyström, Gabriella Sanniti di Baja.) Lecture Notes in Computer Science **1953**, pp. 46–56. Springer.

Kong, T. Y.; Rosenfeld, A.

1989 Digital topology: Introduction and Survey. *Computer vision, graphics, and image processing* **48**, 357–393.

Kong, Yung; Kopperman, Ralph; Meyer, Paul R.

1991 A topological approach to digital topology. *Amer. Math. Monthly* **98**, 901–917.

Kuroš, A. G.

1962 *Lekcii po obščej algebre.* Moscow: F.M.

Rockafellar, R. Tyrrell

1970 *Convex Analysis.* Princeton, NJ: Princeton University Press.

Rosenfeld, Azriel

1979 Digital topology. *Amer. Math. Monthly* **86**, 621–630.

Strömberg, Thomas

1996 The operation of infimal convolution. *Dissertationes Math.* **352**. 58 pp.

Wyse, Frank, et al.

1970 Solution to problem 5712. *Amer. Math. Monthly* **77**, 1119.

Author's address: Uppsala University, Department of Mathematics,
P. O. Box 480, SE-751 06 Uppsala, Sweden.
Telephone: +46 18 4713216 (office); +46 18 300708 (home) Fax: +46 18 4713201
Electronic mail: kiselman@math.uu.se URL: http://www.math.uu.se/~kiselman