2004-10-06

Digital Geometry and Mathematical Morphology

Lecture Notes, Spring Semester 2004

Christer O. Kiselman

Contents:

1. Introduction

Digital geometry is, simply put, the geometry of the computer screen. Mathematical morphology is, in equally simple words, the theory and practice of transformations of sets and functions with an emphasis on their shape. In many cases these transformations have been known for a long time, but they have come into focus for the same reason as digital geometry: the operations can actually be performed on a computer.

1.1. Why digital geometry?

There is a universal answer (*Why not?*) to all such questions, but I shall nevertheless try to give some motivation for this relatively new field and why I think it is worthwhile as a mathematical theory.

Points, straight lines and planes have been studied for well over two millenia, and certain curves, like ellipses and hyperbolas, have been an object for our curiosity for almost as long. Other, less well-known curves, like lemniscates and cardioids, have been studied for several centuries. In the study of these curves we rely very much on the fact that we can draw them on paper. But with the advent of computers we have acquired a new method of drawing pictures. On the computer screen we see images, and the images consist of little picture elements, *pixels*, that the eye puts together to form geometric objects. A straight line is therefore not what Euclid understood by a straight line, but rather a finite collection of dots on the screen, which the eye nevertheless perceives as a connected line segment. Is there a geometry for these images on the screen? The answer is in the affirmative. We shall not be content with the images as more or less accurate approximations of ideal straight lines or curves: we can treat these finite sets of points with the same accuracy as Euclid had in his geometry. This is digital geometry. (Figure 1.)

The field is young in comparison with Euclid's: the notion of a straight line was clarified in 1974 by Azriel Rosenfeld. We can also talk about curves in the digital plane. In fact, we can take any notion in Euclidean geometry, try to translate it to digital geometry, and see if a certain result in Euclidean geometry becomes true in a digital interpretation. The Jordan curve theorem is an instructive example.

For an elementary discussion of digital geometry and in particular of the Jordan curve theorem, see (Kiselman 2003, MS).

The plane may be divided into triangles, rectangles, or hexagons. These are the most common tesselations of the plane. The centers of the pixels form, respectively, a hexagonal, rectangular, or triangular pattern (Figures 2 and 3). In all these cases we may use a pair of integer coordinates to indicate the location of a pixel. This is obvious and easy in the rectangular case, but useful also in the other two cases, although we need to be careful about the metric then. Therefore we often speak of $\mathbf{Z}^2$ as the set of pixels, although, speaking more precisely, a pair of integers $x = (x_1, x_2) \in \mathbf{Z}^2$ is just the address of a pixel.

1.2. Why mathematical morphology?

The field that has become known as mathematical morphology is quite old in a sense; it is about operations on sets and functions that have been around for a long time, but which are now being systematized and studied under a new angle, precisely because it is possible to actually perform operations on the computer and see on the screen what happens.

My personal view is that morphology has its origin in our trying to understand a complicated world. The world is so complex that the human mind—and the human eye—cannot perceive all its minute details, but needs a simplified image, a simplified structure. The need to simplify a complicated object is, in this view of things, the basic impulse behind mathematical morphology, and this is what mathematical morphology does. Related to this is the fact that an image may contain a lot of disturbances, or rather, it almost always does. Therefore, most images need to be tidied up. Hence another need to process images; it is related to the first, for the border line between dirt and other kind of disturbances is not too clear.

Let us think first of Euclidean geometry, and consider cardinalities. The set $\mathbf{N}$ of nonnegative integers is infinite, and its cardinality is denoted by $\operatorname{card}(\mathbf{N}) = \aleph_0$. The set of real numbers $\mathbf{R}$ has the same cardinality as the set of all subsets of $\mathbf{N}$, thus $\operatorname{card}(\mathbf{R}) = 2^{\aleph_0}$. The points in the Euclidean plane have the same cardinality: $\operatorname{card}(\mathbf{R}^2) = \operatorname{card}(\mathbf{R})$. But the set of all subsets of the line or the plane has the larger cardinality $2^{2^{\aleph_0}}$, which is too much for our brains to keep track of. There are simply too many sets in the plane: we need to restrict attention to some not too large subclass of this huge class, a subclass consisting of nice sets. For instance, the set of all disks has a much smaller cardinality, because three numbers suffice to determine a disk in the plane: its radius and the two coordinates of its center. Similarly, four numbers suffice to specify a rectangle $[a_1, b_1] \times [a_2, b_2]$ with sides parallel to the axes; a fifth is needed if we want to rotate it. This leads to the idea of simplifying a general, all too wild set, to some reasonable, more well-behaved set.

It is true that one can think of a Euclidean line as containing only denumerably many points. We can define a line as the set of solutions in $\mathbf{Q}^2$ of an equation $a_1x_1 + a_2x_2 + a_3 = 0$ with integer coefficients. Then two lines which are not parallel intersect in a point with rational coordinates. The cardinality of the set of all subsets of $\mathbf{Q}^2$ is $2^{\aleph_0}$, so there are fewer sets to keep track of than in the real case. However, the conclusion is more or less the same: there are too many subsets of the plane to grasp.

If we consider digital geometry we can make a similar discussion about cardinality. On a computer screen with, say, 1,024 pixels in a horisontal row and 768 pixels in a vertical column there are $1,024 \times 768 = 786,432$ pixels. On such a screen a rectangle with sides parallel to the axes is the Cartesian product

$$R(a,b) = [a_1, b_1]_{\mathbf{Z}} \times [a_2, b_2]_{\mathbf{Z}}$$

of two intervals; we need to consider all pairs of integers (a_j, b_j) such that $0 \leqslant a_1 \leqslant b_1 < 1,024$ and $0 \leqslant a_2 \leqslant b_2 < 768$. We can form

$$\tfrac{1}{2} \cdot 1,024 \cdot 1,025 \cdot \tfrac{1}{2} \cdot 768 \cdot 769 \approx 1.55 \cdot 10^{11}$$

such rectangles. Let us for instance ask if a given subset A is close to a rectangle $R(a,b)$. We can measure the distance between two sets A and B in some way, for instance simply by computing

$$d(A,B) = \operatorname{card}(A \smallsetminus B) + \operatorname{card}(B \smallsetminus A),$$

the number of pixels that belong to one of the sets but not to the other. Then the problem is to compute $d(A, R(a,b))$ for all permissible values of a and b, i.e., in $1.55 \cdot 10^{11}$ cases. At least one of the rectangles is closest to A in this metric. Thus to find the best-fitting rectangle to an arbitrary set is a finite optimization problem ...

But the number of rectangles is very small compared with the number of arbitrary sets. To describe a subset of the screen we need to specify for each pixel whether it belongs to the set or not. This means that there are $2^{786,432} \approx 10^{236,740}$ different subsets of the screen, or binary images. A binary image is a black-and-white image, i.e., we only specify whether a point belongs to the set or to the background. (If we want to consider gray-level images, or color images, the cardinality goes up of course.) The really pure mathematician then says: there are only finitely many binary images. But the number of binary images must be compared with other finite numbers. Some astronomers estimate the mass of the universe at 10^{53} kg, which is 6×10^{79} proton masses or 10^{83} electron masses. One sometimes talks about "astronomical numbers," which Webster defines as "enormously or inconceivably large" numbers. This metaphor has not only faded; it is actually misleading in the world of image analysis.[1]

Thus, although the number of binary images on a computer screen is finite, it is so huge that the conclusion must be the same as in the case of the infinite cardinal $2^{2^{\aleph_0}}$: there are too many; we cannot search through the whole set; we must simplify. This leads, again, to image processing and mathematical morphology, this time of subsets of $\mathbf{Z}^2$, or, a little more generally, of $\mathbf{Z}^n$, the set of all n-tuples of integers.

When we discuss mathematical morphology we want to keep both cases in mind, i.e., both the vector space $\mathbf{R}^n$ of all n-tuples of real numbers (the addresses of points in space) and the digital space $\mathbf{Z}^n$ (the addresses of pixels or voxels). The latter covers the case of rectangular pixels or voxels, but, as already pointed out, also triangular and

[1] In case you think that this conclusion depends on the universe having a rather low density, please calculate the mass of a fictitious universe with a radius of 14×10^9 light years $\approx 1.3 \times 10^{26}$ m and the density of a neutron star, say 10^{17} kg/m^3. The conclusion is the same.

hexagonal pixels in the plane. What is common to $\mathbf{R}^n$ and $\mathbf{Z}^n$ is that they form an abelian group. Therefore we shall always start the discussion assuming that the space, called *image carrier*, is just an abelian group. When necessary we shall then specialize to $\mathbf{R}^n$ and $\mathbf{Z}^n$. In many ways, $\mathbf{R}^n$ will guide us in the perhaps less familiar study of $\mathbf{Z}^n$.

Serra (1982:6—15) lists "four principles of quantification." These are about our ways to gather information about the external world. They apply also, but not exclusively, to image analysis. Let us discuss them briefly in mathematical terms.

Serra's first principle is "compatibility under translation." Mathematically speaking, this means that if we translate the object (i.e., move it some distance without rotating it) and then do something to it, the result should be the same as if we perform the two operations in the other order. For a mapping, this simply means that $f(A+b) = f(A)+b$, which we may express as $f \circ T_b = T_b \circ f$, where $\circ$ denotes *composition of mappings* defined by $(f \circ T_b)(x) = f(T_b(x))$, thus a kind of commutativity, writing T_b for the *translation* translation $T_b(A) = A+b$. We say that f *commutes with translations.* However, when it comes to applications of this idea to images on the computer screen, we are in deep trouble. On a finite screen like $\{x \in \mathbf{Z}^2; 0 \leqslant x_1 < 1,024, 0 \leqslant x_2 < 768\}$ almost nothing can commute with translations. To escape from this difficulty we introduce the ideal, infinite, computer screen with sets of addresses equal to $\mathbf{Z}^2$. The principle is equally useful in $\mathbf{R}^n$ and $\mathbf{Z}^n$.

Serra's second principle is "compatibility under change of scale." For a mapping this means that it commutes with *homotheties* (or *dilatations*), i.e., mappings of the form $x \mapsto \lambda x$. So $f(\lambda A) = \lambda f(A)$ for (say) positive number λ. This is not problematic if we are in $\mathbf{R}^n$, but it certainly is if we are in $\mathbf{Z}^n$. This principle will therefore need to be suitably interpreted in $\mathbf{Z}^n$.

The third principle is that of "local knowledge." This principle says that in order to know some bounded part of $f(A)$, we shall not need to know all of A, only some bounded part of A. Mathematically speaking: for every bounded set Y, there exists a bounded set Z such that $f(A \cap Z) \cap Y = f(A) \cap Y$. To know the result $f(A)$ in an arbitrary bounded set Y, we need not know all of A; it is enough to know how A looks in some bounded set Z, maybe a little larger than Y. The principle of local knowledge shall therefore not be understood in the topological sense: the key notion is boundedness.

Serra's fourth principle of quantification is that of "semicontinuity." It means that if a decreasing sequence (A_j) of closed sets tends to a limit A, thus $A = \bigcap A_j$, then $f(A_j)$ tends to $f(A)$. Thus if A_j is close to A in some sense and A_j contains A, then $f(A_j)$ must be close to $f(A)$. To express this property as semicontinuity, one must define a topology.

2. Preordered sets

2.1. Preorders and orders

For the morphological operations on the computer screen we need to consider sets of sets. The set of all subsets of a given set is ordered by the inclusion relation, which is an example of an abstract order relation. It is therefore convenient to introduce notions that will be useful in the general theory of order relations. In this chapter we shall do so.

Let us define first the notion of preorder, which is even more fundamental than that of order.

Definition 2.1.1. *A* preorder *in a given set X is a relation (a subset of X^2) which is reflexive and transitive.*

The definition means, if we denote the relation by $\leqslant$, that for all $x, y, z \in X$,

$$x \leqslant x, \tag{2.1.1}$$

and

$$x \leqslant y \text{ and } y \leqslant z \text{ implies } x \leqslant z. \tag{2.1.2}$$

Definition 2.1.2. *An* order *is a preorder which is antisymmetric.*

This means that the relation shall satisfy

$$x \leqslant y \text{ and } y \leqslant x \text{ implies } x = y. \tag{2.1.3}$$

(Sometimes one calls such a relation a *partial order.*)

A *preordered set* is a set together with a preorder; an *ordered set* a set together with an order.

A basic example of an ordered set is the set *power set* $\mathscr{P}(W)$ 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 $A, B \in \mathscr{P}(W)$.

Suppose that we have two preorders defined in a set X; denote them by $\leqslant$ and $\preccurlyeq$. The preorder $\leqslant$ is said to be *finer* than the preorder $\preccurlyeq$, and $\preccurlyeq$ is said to be *coarser* than $\leqslant$, if $x \leqslant y$ implies $x \preccurlyeq y$ for all x, y.

There is a finest preorder in a set, viz. when we define $x \leqslant y$ to mean that $x = y$. This preorder is of course an order; let us call it the *discrete order.* There is also a coursest preorder in any set X, when we declare that $x \leqslant y$ for all $x, y \in X$. Let us call this the *chaotic preorder.* The set of all preorders on any set is thus an ordered set with a largest and a smallest element.

Given any preorder we define $[a, \rightarrow[$ as the set of all x such that $a \leqslant x$ and $]\leftarrow, a]$ as the set of all x such that $x \leqslant a$. The sets $[a, \rightarrow[$ and $]\leftarrow, a]$ determine a for all $a \in X$ if and only if $\leqslant$ is an order.

Definition 2.1.3. *An* equivalence relation *is a preorder which is symmetric.*

Given a preorder $\leqslant$ in X, we can introduce an equivalence relation $\simeq$ in X by defining $x \simeq y$ to mean that $x \leqslant y$ and $y \leqslant x$. If $\leqslant$ is an order, then $\simeq$ is just equality. We can form the quotient space $X/\simeq$ of all equivalence classes of X modulo $\simeq$. The equivalent classes are just

$$[a, \rightarrow[\ \cap \ ]\leftarrow, a] = \{x \in X; a \leqslant x \leqslant a\}, \qquad a \in X.$$

In this quotient space, $\leqslant$ induces an order (exercise 2.1).

In preordered spaces the increasing mappings are of importance:

Definition 2.1.4. *If $f\colon X \to Y$ is a mapping from a preordered set X to a preordered set Y, then we say that f is* increasing[2] *if*

$$\text{for all } x, x' \in X, \text{ the relation } x \leqslant_X x' \text{ implies } f(x) \leqslant_Y f(x'). \tag{2.1.4}$$

We shall write $\mathsf{Incr}(X, Y)$ for the set of all increasing mappings $X \to Y$. We may think of $\mathsf{Incr}(X, Y)$ as an analogue of the linear mappings from a vector space into another.

A preorder $\leqslant$ is finer than another preorder $\preccurlyeq$ if and only if the identity mapping $(X, \leqslant) \to (X, \preccurlyeq)$ is increasing.

Note that if X has the discrete order, where $x \leqslant y$ means $x = y$, then $\mathsf{Incr}(X, Y)$ consists of all mappings $X \to Y$; $\mathsf{Incr}(X, Y) = Y^X$. The conclusion is the same if Y is equipped with the chaotic preorder. If on the other hand Y has the discrete order, then $\mathsf{Incr}(X, Y)$ consists of those mappings $X \to Y$ that are constant on any chain of comparable elements, thus $f(x_j)$ is constant if $x_1 \leqslant x_2 \geqslant x_3 \leqslant \cdots \leqslant x_{2n}$. Let us say that a preordered set X is *connected* if, given any elements $a, b \in X$, there exists a finite chain of elements $x_1 \leqslant x_2 \geqslant x_3 \leqslant \cdots \leqslant x_{2n}$ passing through a and b. Then any mapping $f \in \mathsf{Incr}(X, Y)$ is constant if X is connected and Y has the discrete order. The same conlusion holds if X has the chaotic preorder and Y is ordered.

Another interesting property is obtained when we turn around the implication sign in (2.1.4).

Definition 2.1.5. *Let us agree to call a mapping $f\colon X \to Y$* coincreasing *if it satisfies*

$$\text{for all } x, x' \in X, \text{ the relation } f(x) \leqslant_Y f(x') \text{ implies } x \leqslant_X x'. \tag{2.1.5}$$

Increasing does not imply coincreasing; coincreasing does not imply increasing; cf. exercise 2.2. If X is ordered, then a mapping is coincreasing if and only if it is injective and the mapping $f(x) \mapsto x$ is increasing.

2.2. Closure operators

Definition 2.2.1. *A* closure operator[3] *(or* closing[4]*) in an ordered set X is a mapping $X \ni x \mapsto \overline{x} \in X$ which is* increasing, idempotent, *and* extensive[5] *(or* expanding*); in other words, which satisfies the following three conditions for all $x, y \in X$:*

$$x \leqslant y \text{ implies } \overline{x} \leqslant \overline{y}; \tag{2.2.1}$$

$$\overline{\overline{x}} = \overline{x}; \tag{2.2.2}$$

$$x \leqslant \overline{x}; \tag{2.2.3}$$

[2] Sometimes these mappings are called *isotone* (Birkhoff 1948:49) or *order preserving.*

[3] Ore (1944:494) and Birkhoff (1948:49) used the term *closure operation*; the latter attributed the concept to Moore (1910:53—80). However, although Moore wrote about *closure* and *closure properties*, he did not give a clearcut definition. Everett (1944) used *closure operator.* Dubreil & Dubreil-Jacotin (1964:9, 177) calls the operator *fermeture de Moore.*

[4] This term seems to have been introduced by Matheron (1975:18). In Matheron (1967:18) he used the French term *fermeture* for the notion.

[5] Birkhoff (1948:49).

When checking (2.2.2) it is of course enough to prove that $\overline{\overline{x}} \leqslant \overline{x}$ if we have already proved (2.2.3).

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).

Sometimes we shall consider operators that are defined only on a subset of X.

Let $f\colon Y \to X$ be a mapping, where $Y \subset X$. Then it is not obvious what idempotency shall mean, for the composition $f(f(y))$ need not be defined. We solve this problem by a modified definition; see Kiselman (1969:336).[6]

Definition 2.2.2. *We shall say that $\gamma\colon Y \to X$, where Y is a subset of an ordered set X, is a* closure operator *if it satisfies the following two conditions for all $x, y \in Y$:*

$$x \leqslant \gamma(x); \tag{2.2.4}$$

$$x \leqslant \gamma(y) \text{ implies } \gamma(x) \leqslant \gamma(y). \tag{2.2.5}$$

It follows from this that γ is increasing, and that $\gamma(\gamma(x)) = x$ for every $y \in Y$ such that $\gamma(y)$ happens to belong to Y. It is also clear that if $Y = X$, then the new definition agrees with Definition 2.2.1.

Any closure operator γ in the sense of Definition 2.2.2 can be extended to an idempotent operator $\gamma_1\colon Z \to X$ by taking $Z = Y \cup \operatorname{im}\gamma$ and defining $\gamma_1(x) = x$ when $x \in Z \smallsetminus Y$.

A basic example of a closure operator is 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 $\mathscr{P}(W)$ defines a topology in W if and only if it satisfies, in addition to (2.2.1), (2.2.2), (2.2.3) above, two extra conditions, viz.

$$\overline{\emptyset} = \emptyset \text{ and } \overline{A \cup B} = \overline{A} \cup \overline{B} \text{ for all } A, B \subset W, \tag{2.2.6}$$

where $\emptyset$ denotes *the empty set*, the set with no elements.

Another closure operator of great importance is the operator which associates to a set A in $\mathbf{R}^n$ its convex hull, the smallest convex set containing the given set, denoted by $\operatorname{cvx} A$. The composition $A \mapsto \overline{\operatorname{cvx} A}$ is a closure operator, whereas the composition in the other order, $A \mapsto \operatorname{cvx}\bigl(\overline{A}\bigr)$ is not idempotent (exercise 2.3). We see that the composition of two closure operators is sometimes, but not always, a closure operator.

In both these examples X is the power set $\mathscr{P}(\mathscr{W})$ of some set W, and the closure operator is given as an intersection:

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

In Chapter 7 we shall return to these constructions.

[6] Matheron (1975:186) required instead that $f(y)$ belong to Y whenever $y \in Y$.

Dual to the notion of closure operator is the notion of *opening.*

Definition 2.2.3. *A mapping $X \ni x \mapsto x^\circ \in X$ is said to be an* opening[7] *if it is antiextensive, increasing, and idempotent; in other words, if it satisfies the following three conditions for all $x, y \in X$:*

$$x^\circ \leqslant x; \tag{2.2.7}$$

$$x \leqslant y \text{ implies } x^\circ \leqslant y^\circ; \tag{2.2.8}$$

$$(x^\circ)^\circ = x^\circ. \tag{2.2.9}$$

Just as for closure operators, we can extend this definition to a more general setting:

Definition 2.2.4. *Let $\eta\colon Y \to X$ be a mapping of a subset Y into X. We shall say that η is an* opening *if it satisfies the following two conditions for all $x, y \in Y$:*

$$\eta(x) \leqslant x; \tag{2.2.10}$$

$$\eta(x) \leqslant y \text{ implies } \eta(x) \leqslant \eta(y). \tag{2.2.11}$$

2.3. Exercises

2.1. Verify that any preorder in a set X induces an order in the quotient $X/\simeq$, where $x \simeq y$ denotes the equivalence relation $x \leqslant y \leqslant x$.

2.2. Consider the following four properties of a mapping $f\colon X \to Y$ between two preordered sets.
(A) f is increasing: $x \leqslant y \Rightarrow f(x) \leqslant f(y)$;
(B) f is strictly increasing: $x < y \Rightarrow f(x) < f(y)$, where $x < y$ means that $x \leqslant y$ and $x \neq y$;
(C) f is injective: $f(x) = f(y) \Rightarrow x = y$;
(D) f is coincreasing: $f(x) \leqslant f(y) \Rightarrow x \leqslant y$.

Prove that (B) implies (A); that (A)&(C) implies (B); and that, if X is ordered, (D) implies (C).

Prove by examples that the implications (B) ⇒ (C); (B)&(C) ⇒ (D); (D) ⇒ (A) do not hold, not even for mappings $f\colon X \to X$.

If we assume that X is ordered and that (A) holds, we thus have (D) ⇒ (C) and (C) ⇒ (B), but the converse implications do not hold.

2.3. Let f denote the mapping $\mathscr{P}(\mathbf{R}^2) \ni A \mapsto \overline{A} \in \mathscr{P}(\mathbf{R}^2)$ of taking the topological closure of a set A and g the mapping $A \mapsto \mathrm{cvx}A$ of taking the convex hull of A. Are $f \circ g$ and $g \circ f$ closure operators? Prove or disprove.

3. Morphological operations on sets and functions

3.1. Dilations and erosions

Our purpose is to describe morphological operations in $\mathbf{Z}^2$ as well as in $\mathbf{R}^2$ and more generally in $\mathbf{Z}^n$ and $\mathbf{R}^n$. It will therefore be an advantage if we can choose a common

[7] This term seems to have been introduced by Matheron (1975:18). In his earlier book (1967:18) he used the French term *ouverture.* Tucker (1936:94) used the term *aperture* for the dual of closure.

framework for these two cases. Since $\mathbf{Z}^2$ is not a vector space, we cannot assume that we live in vector space. What $\mathbf{Z}^n$ and $\mathbf{R}^n$ have in common is that they are abelian groups. It turns out that this concept is rich enough for a theory of morphological operations to be set up.

In an abelian group G we have an associative and commutative operation, written as addition, and a *neutral element*, denoted by 0 (i.e., such that $x + 0 = x$ for all $x \in G$); finally every element x has an inverse, written $-x$ and satisfying $x + (-x) = 0$. This generality causes no problem in the definitions and proofs. However, the reader can think of the special abelian groups $\mathbf{Z}^2$ and $\mathbf{R}^2$ all the time.

For some definitions and results it is not necessary to assume the group to be commutative; sometimes it is not even necessary to assume that we have a group: a semigroup will suffice. A *semigroup* is a set together with an associative operation, written as juxtaposition. We do not assume commutativity, nor the existence of a neutral element. If there is one, it will be denoted by 1. The reason for assuming so little is not generality in the first place but a kind of Occam's razor: by not assuming too much we make the constructions more transparent and clarify the dependence of a result on the hypotheses.

Definition 3.1.1. *Let A and B be subsets of a semigroup G. Then we define their* product *as the set*

$$AB = \{xy; x \in A, y \in B\}. \tag{3.1.1}$$

If the semigroup is commutative, we shall write + for the operation and define the Minkowski sum *of two sets A and B as*

$$A + B = \{x + y; x \in A, y \in B\}. \tag{3.1.2}$$

The operation makes the power set $\mathscr{P}(G)$ of all subsets of G into a semigroup. In that semigroup the empty set $\emptyset$ is a *zero*: $A\emptyset = \emptyset A = \emptyset$; and $\{1\}$ is a neutral element if G happens to have a neutral element 1: $A\{1\} = \{1\}A = A$.

In the commutative case the operation $(A, B) \mapsto A+B$ is called *Minkowski addition.* We have $A + \emptyset = \emptyset$ and $A + \{0\} = A$ if G has a neutral element 0.

If B is finite, as is often the case in $\mathbf{Z}^n$, only finitely many checks are needed to decide whether a point x belongs to $A + B$: we check whether $x - b$ belongs to A for some $b \in B$. Therefore dilation by a finite set satisfies Serra's third principle (see the introduction, section 1.2).

If A consists of only one point x we shall write $AB = \{x\}B = xB = S_x(B)$, where the last equality is a definition of an operation $S_x \colon \mathscr{P}(G) \to \mathscr{P}(G)$, called *left translation by x*; all points are moved by a fixed amount. Similarly we define *right translation by y* as the operation T_y, $T_y(A) = A\{y\} = Ay$. In the commutative case the two translations are equal, and we write $S_y(A) = y + A = A + y = T_y(A)$. If G is a group, S_x and T_y have inverses, viz. $(S_x)^{-1} = S_{x^{-1}}$ and $(T_y)^{-1} = T_{y^{-1}}$ (in the commutative case $(S_x)^{-1} = (T_x)^{-1} = S_{-x} = T_{-x}$).

Using the concept of translation we see that

$$AB = \bigcup_{y \in B} (Ay) = \bigcup_{x \in A} (xB), \qquad A + B = \bigcup_{y \in B} (A + y) = \bigcup_{x \in A} (x + B). \tag{3.1.3}$$

Thus the product (sum) is a union of translates of A, and equally a union of translates of B. We can view it in either way: we move A around in a manner determined by B or vice versa. Often A is a complicated set and B is a simple, small set, called the *structural element*. Then we say that AB (or $A+B$) is A *dilated* by B.[8] Think of $G=\mathbf{R}^2$ or $\mathbf{Z}^2$ and with B equal to a small square. We fix B and apply the operation $A\mapsto A+B$ to a lot of sets A:

Definition 3.1.2. *Let G be a semigroup and let B be a subset of G. We define two mappings $\delta_B, \varepsilon_B\colon \mathscr{P}(G)\to\mathscr{P}(G)$ as*

$$\delta_B(A)=AB, \qquad \varepsilon_B(A)=\{x; xB\subset A\}. \tag{3.1.4}$$

The first is called dilation[9] *by B. The second is called* erosion by B.

To stress the similarity between the two definitions we may rewrite them as

$$\delta_B(A)=\bigcup_x\big(\{x\}B; \{x\}\subset A\big), \qquad \varepsilon_B(A)=\bigcup_x\big(\{x\}; \{x\}B\subset A\big). \tag{3.1.5}$$

In the commutative case these formulas take the form

$$\delta_B(A)=A+B, \qquad \varepsilon_B(A)=\{x; x+B\subset A\}, \tag{3.1.6}$$

and

$$\delta_B(A)=\bigcup_x\big(\{x\}+B; \{x\}\subset A\big), \qquad \varepsilon_B(A)=\bigcup_x\big(\{x\}; \{x\}+B\subset A\big). \tag{3.1.7}$$

We immediately note an important relation between the two operations.

Proposition 3.1.3. *Let A, B and C be three subsets of a semigroup. Then the following three statements are equivalent.*
(a) *$AB\subset C$ (in the commutative case $A+B\subset C$);*
(b) *$\delta_B(A)\subset C$;*
(c) *$A\subset\varepsilon_B(C)$.*

Proof. That (a) and (b) are equivalent is obvious from the definition of dilation. The definition of erosion can be written

$$x\in\varepsilon_B(C)\Leftrightarrow xB\subset C.$$

If (a) holds, the statement to the right holds for all $x\in A$, hence $A\subset\varepsilon_B(C)$, i.e., (c) is true. Conversely, if (c) holds, then the statement to the left holds for all $x\in A$, hence $AB\subset C$.

We note that dilation commutes with left translation:

$$S_x(\delta_B(A))=x(AB)=(xA)B=\delta_B(S_x(A)), \qquad A\in\mathscr{P}(G),$$

[8] Sometimes $A+B$ is written $A\oplus B$; see Matheron (1975:16), Serra (1982, 2001) and Gonzalez & Woods (1993:519); note however that the sign $\oplus$ is used for the direct sum. There is no risk of misunderstanding $A+B$.

[9] This seems to be the most common term today; Matheron (1975:17) calls it *dilatation*, a term which is often reserved for mappings $x\mapsto\lambda x$.

which may be written as $S_x \circ \delta_B = \delta_B \circ S_x$. Similarly, if G is a group, erosion commutes with left translation:

$$S_z(\varepsilon_B(A)) = \{zx; xB \subset A\} = \{y; z^{-1}yB \subset A\} = \{y; yB \subset zA\} = \varepsilon_B(S_z(A)),$$

thus $S_z \circ \varepsilon_B = \varepsilon_B \circ S_z$.

Dilation commutes also with the formation of unions:

$$\bigcup(A_jB) = \left(\bigcup A_j\right)B, \text{ in the commutative case } \bigcup(A_j + B) = \left(\bigcup A_j\right) + B, \tag{3.1.8}$$

where $(A_j)_{j\in J}$ is a finite or infinite family of subsets of an arbitrary semigroup G. There is a converse to this statement:

Proposition 3.1.4. *Let G be a semigroup with neutral element 1. Any mapping $f\colon \mathscr{P}(G) \to \mathscr{P}(G)$ which commutes with left translations and the formation of infinite unions is a dilation, in fact by $f(\{1\})$.*

Proof. We write $A = \bigcup_{x\in A}\{x\} = \bigcup\{x\}$ so that

$$f(A) = f\left(\bigcup\{x\}\right) = \bigcup f(\{x\}) = \bigcup f(S_x\{1\}) = \bigcup S_x(f(\{1\})) = Af(\{1\}) = \delta_{f(\{1\})}(A).$$

Similarly, erosion commutes with the formation of arbitrary intersections:

$$\varepsilon_B\left(\bigcap A_j\right) = \{x; xB \subset \bigcap A_j\} = \bigcap\{x; xB \subset A_j\} = \bigcap \varepsilon_B(A_j). \tag{3.1.9}$$

A converse to this statement will be established later; see Proposition 3.1.7.

If we assume that G is not only a semigroup but a group, we can use inverses to define sets. To any given subset A of a group G we define its *opposite set* as $\check{A} = \{x^{-1}; x \in A\}$; in the commutative case $\check{A} = \{-x; x \in A\}$. Then we can also define *Minkowski subtraction* by the formula

$$A - B = A + \check{B} = \{x - y; x \in A, y \in B\}.$$

Note that $\check{B} = \{0\} - B = 0 - B$; we may also write it as $-B$. We should not mix this up with the *set-theoretical difference*:

$$A \smallsetminus B = A \cap \complement B = A \cap B^c = \{x \in A; x \notin B\},$$

where we have written the complement of B in two different ways: $\complement B = B^c$; both are quite usual. We now introduce another minus sign in an abelian group, written $\ominus$:[10]

$$A \ominus B = \varepsilon_B(A) = \{x; x + B \subset A\}. \tag{3.1.10}$$

Note that $A \ominus \emptyset = G$, that $A \ominus G = \emptyset$ provided $A \neq G$, and that $A \ominus \{0\} = A$ if G has a neutral element 0.

To illustrate the difference in meaning between these three minus signs, note that $A \smallsetminus A = \emptyset$ for all sets A; that $A - A \neq \emptyset$ if A is a nonempty subset of an abelian group; and that, finally, $A \ominus A$ is a subsemigroup if A is a subset of an abelian group

[10] Gonzalez & Woods (1993:521) use the sign $\ominus$ with this definition; Serra (1982:43) on the other hand defines $A \ominus B$ as $\bigcap_{y\in B}(A + y) = \varepsilon_{\check{B}}(A)$.

G, cf. Theorem 3.1.5 below. If A is a bounded subset of $\mathbf{R}^n$ or $\mathbf{Z}^n$ for example, then $A \ominus A = \{0\}$.

In a group, erosion is dual to dilation in a natural sense. Equivalently, an erosion of a set A is an intersection of translates of A.

To any mapping $f\colon \mathscr{P}(W_1) \to \mathscr{P}(W_2)$ we may define its *dual mapping* $f^{\mathbf{d}}$ by

$$f^{\mathbf{d}}(A) = \complement f\big(\complement A\big)\,, \qquad A \in \mathscr{P}(W_1).$$

We shall now see that, in a group, erosion by B is dual to dilation by $\check{B}$.

Theorem 3.1.5. *Let G be a group and B a subset of G. Then*

$$\varepsilon_B(A) = \bigcap_{y\in B} Ay^{-1} = \complement \bigcup_{y\in B} \complement\left(Ay^{-1}\right) = \complement \bigcup_{y\in B} \left(\left(\complement A\right) y^{-1}\right) = \complement\delta_{\check{B}}\left(\complement A\right). \tag{3.1.11}$$

With additive notation this becomes

$$\varepsilon_B(A) = \bigcap_{y\in B} (A-y) = \complement \bigcup_{y\in B} \complement\,(A-y) = \complement \bigcup_{y\in B} \left(\left(\complement A\right) - y\right) = \complement\delta_{\check{B}}\left(\complement A\right). \tag{3.1.12}$$

In the commutative case we can write

$$A \ominus B = \varepsilon_B(A) = \{x; x + B \subset A\} = \complement\left(\complement A + \check{B}\right). \tag{3.1.13}$$

Proof. To prove the first equality, note that x belongs to $\varepsilon_B(A)$ if and only if $xy \in A$ for all $y \in B$, which is equivalent to $x \in Ay^{-1}$ for all $y \in B$, i.e., to x belonging to $\bigcap_{y\in B} Ay^{-1}$. The second equality is just one of De Morgan's laws.

The third equality depends on the fact that the formation of complement commutes with right translation; see the next lemma. The fourth and last equality follows from the definition of $\delta_{\check{B}}$. We are done.

In the proof of the theorem we needed the following result.

Lemma 3.1.6. *Let G be a group. Then the formation of complement commutes with left and right translation:*

$$\complement(zA) = z\big(\complement A\big) \ \text{ and } \ \complement(Az) = \big(\complement A\big)z, \qquad A \in \mathscr{P}(G), \quad z \in G,$$

briefly $\complement \circ S_z = S_z \circ \complement$ and $\complement \circ T_z = T_z \circ \complement$.

Proof. If we analyze the relations $x \in \complement(zA)$ and $x \in z\big(\complement A\big)$ we find that they are both equivalent to $z^{-1}x \notin A$. Similarly, $y \in \complement(Az)$ and $y \in \big(\complement A\big)z$ are equivalent to $yz^{-1} \notin A$.

Thus Theorem 3.1.5 says that

$$\varepsilon_B = (\delta_{\check{B}})^{\mathbf{d}} = \complement \circ \delta_{\check{B}} \circ \complement,$$

from which it follows that $\delta_B = (\varepsilon_{\check{B}})^{\mathbf{d}}$.

From Proposition 3.1.4 and Theorem 3.1.5 we now deduce a result on erosions using this duality.

Proposition 3.1.7. *Let G be a group. Any mapping $g\colon \mathscr{P}(G) \to \mathscr{P}(G)$ which commutes with left translations and the formation of infinite intersections is an erosion, viz. by $\complement\left(g\left(\complement\{1\}\right)\right)\check{}$.*

3.2. Infimal convolution

The product AB of two subsets can be put into the wider framework of infimal convolution. Given two functions $f, g\colon G \to [-\infty, +\infty]$ defined on a semigroup G and with values in the *extended real line* $[-\infty, +\infty] = \mathbf{R} \cup \{-\infty, +\infty\}$, we define a new function $h = f \mathbin{\square} g$, called the *infimal convolution* of f and g, as

$$(f \mathbin{\square} g)(z) = h(z) = \inf_{x,y\in G} \bigl(f(x) \mathbin{\dot{+}} g(y); xy = z\bigr), \qquad z \in G. \tag{3.2.1}$$

The infimum is taken over all elements $x, y \in G$ such that their product is z, the argument of h. If there are no such elements, the infimum is plus infinity; that element is by definition the infimum over the empty set. There is a complication if f takes the value $+\infty$ at x and g takes the value $-\infty$ at y. We resolve this conflict by declaring that $+\infty$ shall win. So $s \mathbin{\dot{+}} t$ is the usual sum if s and t are real numbers; if only one is infinite or both are infinite of the same sign, the sum takes that value; if s and t are infinite of opposite signs, we define the sum to be $+\infty$. In this way, this operation, called *upper addition*, becomes an upper semicontinuous mapping from $[-\infty, +\infty]^2$ into $[-\infty, +\infty]$. It is easy to check that it is associative. Similarly we define *lower addition*, $s \mathbin{\underset{\cdot}{+}} t = -((-s) \mathbin{\dot{+}} (-t))$; here minus infinity wins. In my experience, using upper and lower addition is the most convenient method to calculate with the two infinities. This will be apparent, I hope, when we want to show that infimal convolution is associative.

For surveys of the properties of infimal convolution we refer to Moreau (1970), Rockafellar (1970), or Strömberg (1996).

The points where f or g takes the value $+\infty$ play no role in the formation of the infimum: the definition of upper addition guarantees this. Removing these points therefore yields an equivalent definition:

$$(f \mathbin{\square} g)(z) = \inf_{x,y\in G} \left(f(x) + g(y); xy = z, f(x) < +\infty, g(y) < +\infty\right), \qquad z \in G. \tag{3.2.2}$$

The *effective domain*, written $\operatorname{dom} f$, of a function $f\colon X \to [-\infty, +\infty]$ defined on an arbitrary set X is the set where it is strictly less than plus infinity:

$$\operatorname{dom} f = \{x \in X; f(x) < +\infty\}. \tag{3.2.3}$$

With this notation we can write (3.2.2) as

$$(f \mathbin{\square} g)(z) = \inf_{\substack{x\in\operatorname{dom} f\\ y\in\operatorname{dom} g\\ xy=z}} \bigl(f(x) + g(y)\bigr), \qquad z \in G. \tag{3.2.4}$$

Here there is no doubt about the meaning of the sum; the disadvantage with this approach is that we have to remember, each time we take an infimum, over which set the variables range.

Intuitively, plus infinity corresponds to vacuum and $-\infty$ to an infinitely dense neutron star. This may appear to be upside down. However, we should think of the density as $e^{-f(x)}$, and then of course $e^{-(+\infty)} = 0$. Infimal convolution is related to a kind of supremal convolution of the functions e^{-f}, e^{-g}, viz.

$$\sup_y \left[e^{-f(y)} e^{-g(x-y)}\right] = e^{-(f \mathbin{\square} g)(x)}.$$

The supremum is often comparable to integration in $\mathbf{R}^n$, which means that we sometimes have a remarkably good approximation

$$e^{(f \,\square\, g)(x)} = \sup_{y\in\mathbf{R}^n} \left[e^{-f(y)}e^{-g(x-y)}\right] \approx \int_{\mathbf{R}^n} e^{-f(y)}e^{-g(x-y)}dy = \left(e^{-f} * e^{-g}\right)(x), \qquad x \in \mathbf{R}^n,$$

where the asterisk denotes usual convolution, which is defined by the integral

$$(3.2.5) \qquad (F * G)(x) = \int_{\mathbf{R}^n} F(y)G(x-y)dy, \qquad x \in \mathbf{R}^n.$$

Proposition 3.2.1. *Infimal convolution is associative:* $(f_1 \,\square\, f_2) \,\square\, f_3 = f_1 \,\square\, (f_2 \,\square\, f_3)$.

Proof. We start calculating:

$$\big((f_1 \,\square\, f_2) \,\square\, f_3\big)(z) = \inf_{\substack{y,x_3\\ yx_3=z}} \left(\left[\inf_{\substack{x_1,x_2\\ x_1x_2=y}} \big(f_1(x_1) \dotplus f_2(x_2)\big)\right] \dotplus f_3(x_3)\right).$$

According to the following lemma this expression is equal to

$$\inf_{\substack{y,x_3\\ yx_3=z}} \inf_{\substack{x_1,x_2\\ x_1x_2=y}} \big(\left[f_1(x_1) \dotplus f_2(x_2)\right] \dotplus f_3(x_3)\big) = \inf_{\substack{x_1,x_2,x_3\\ x_1x_2x_3=z}} \big(f_1(x_1) \dotplus f_2(x_2) \dotplus f_3(x_3)\big).$$

A similar calculation shows that also $(f_1 \,\square\, (f_2 \,\square\, f_3))(z)$ can be transformed to the last expression. This proves associativity.

At a point in the proof above we needed the following result.

Lemma 3.2.2. *For any element* $c \in [-\infty, +\infty]$ *and any function* $f \colon X \to [-\infty, +\infty]$ *defined on an arbitrary set* X *we have*

$$\inf_{x\in X} (c \dotplus f(x)) = c \dotplus \inf_{x\in X} f(x).$$

Proof. We just need to check all possibilities where our intuition is less reliable than usual, i.e., when $c = \pm\infty$ or X is empty.

Note that there are no exceptions to this formula. As a nasty little exercise, try to find the exact conditions under which the equality $\sup_x(c \dotplus f(x)) = c \dotplus \sup_x f(x)$ holds.

If G is a group, we know that $xy = z$ if and only if $y = x^{-1}z$, which in turn is equivalent to $x = zy^{-1}$, so the definition of $f \,\square\, g$ can be written

$$(f \,\square\, g)(z) = \inf_{x\in G} \big(f(x) \dotplus g(x^{-1}z)\big) = \inf_{y\in G} \big(f(zy^{-1}) \dotplus g(y)\big), \qquad z \in G.$$

In the case of an abelian group this formula reads

$$(f \,\square\, g)(z) = \inf_{x\in G} \big(f(x) \dotplus g(z-x)\big) = \inf_{y\in G} \big(f(z-y) \dotplus g(y)\big), \qquad z \in G.$$

Now why is infimal convolution more general that the formation of products of sets or Minkowski addition? This is because of the formula

$$(3.2.6) \qquad \operatorname{dom}(f \,\square\, g) = \operatorname{dom} f + \operatorname{dom} g,$$

which is easily proved. A special case of this formula is obtained when we consider indicator functions.

To any subset A of a set X we define its *indicator function* i_A, which is simply defined as $i_A(x) = 0$ when $x \in A$ and $i_A(x) = +\infty$ when $x \notin A$. It is related to the *characteristic function* χ_A of A by the formula $\chi_A = \exp(-i_A)$.

It is clear that $\operatorname{dom} i_A = A$. We then have $i_A \mathbin{\square} i_B = i_{AB}$ for all subsets A, B of a semigroup G, or, in the commutative case, $i_A \mathbin{\square} i_B = i_{A+B}$. Hence the Minkowski sum may be defined in terms of infimal convolution as $A + B = \operatorname{dom}(i_A \mathbin{\square} i_B)$.

If G has a neutral element 1, then $i_{\{1\}}$ is a neutral element for infimal convolution. More generally, if $f(x) = 0$ when $x \in A$ and $f(x) \geqslant 1$ otherwise and similarly with g and B, then $(f \mathbin{\square} g)(z) = 0$ when $z \in AB$ and $(f \mathbin{\square} g)(z) \geqslant 1$ when $z \notin AB$.

Thus multiplication of sets and Minkowski addition can be expressed in terms of infimal convolution. But we can go also in the other direction and express any infimal convolution as a product of sets, albeit at the expense of adding one more dimension. This is done using the notion of strict epigraph.

Let us first define the *epigraph* of a function $f\colon X \to [-\infty, +\infty]$ defined on an arbitrary set X as

$$\operatorname{epi} f = \{(x,t) \in X \times \mathbf{R}; f(x) \leqslant t\}, \tag{3.2.7}$$

and the *strict epigraph* using instead strict inequality:

$$\operatorname{epi_s} f = \{(x,t) \in X \times \mathbf{R}; f(x) < t\}. \tag{3.2.8}$$

If $X = G$ is a semigroup, then we make $G \times \mathbf{R}$ into a semigroup by defining $(x,s)(y,t) = (xy, s+t)$. It is not difficult to show that

$$\operatorname{epi_s}(f \mathbin{\square} g) = (\operatorname{epi_s} f)(\operatorname{epi_s} g). \tag{3.2.9}$$

This means that the function $f \mathbin{\square} g$ can be defined as the function whose strict epigraph is the product $(\operatorname{epi_s} f)(\operatorname{epi_s} g)$. With the additive notation we have of course

$$\operatorname{epi_s}(f \mathbin{\square} g) = \operatorname{epi_s} f + \operatorname{epi_s} g. \tag{3.2.10}$$

For the epigraph we always have

$$\operatorname{epi}(f \mathbin{\square} g) \supset \operatorname{epi} f + \operatorname{epi} g,$$

where the inclusion relation may be strict (find examples).

3.3. Exercises

3.1. Dilate and erode A by B when

(a) A is a disk in $\mathbf{R}^2$, B a pair of points. (Here the word *disk* shall be understood with regard to one of the norms $\|\cdot\|_p$, $p = 1, 2, \infty$.)

(b) A is an open disk in $\mathbf{R}^2$ of radius R, B a closed disk of radius r.

(c) B is a translate of A.

(d) B is a translate of $\check{A}$.

3.2. We know that $\delta_B(A) \subset C$ if and only if $A \subset \varepsilon_B(C)$, in other words that $A + B \subset C$ if and only if $A \subset C \ominus B$.

(a) Give examples to show that we do not have an equivalence $\delta_B(A) = C$ if and only if $A = \varepsilon_B(C)$. Is one of the implications true? However, the equivalence holds for certain families of sets ...

(b) Show by examples that it is not true that $\varepsilon_B(A) \subset C$ if and only if $A \subset \delta_B(C)$. Is one of the implications true?

3.3. Prove that, in an abelian group, erosion by B satisfies $\varepsilon_B(A) \subset A$ for all sets A if and only if $0 \in B$.

3.4. Calculate the Minkowski sum of two segments $[a, b] = \{(1-t)a + tb; 0 \leqslant t \leqslant 1\}$ and $[c, d]$ in $\mathbf{R}^2$, where a, b, c, d are four arbitrary points in $\mathbf{R}^2$.

3.5. Calculate in $\mathbf{R}^2$ the sum of three arbitrary segments.

3.6. Calculate in $\mathbf{R}^3$ the sum of three orthogonal segments.

3.7. Calculate in $\mathbf{R}^3$ the Minkowski sum of four segments, three of which are contained in a plane while the fourth is not.

3.8. (a) Calculate in $\mathbf{R}^3$ the sum of the four segments $[(0, 0, 0), (1, 0, 0)]$, $[(0, 0, 0), (0, 1, 0)]$, $[(0, 0, 0), (0, 0, 1)]$, and $[(0, 0, 0), (1, 1, 1)]$. The result is a polyhedron. How many vertices, edges, and faces does it have?

(b) Calculate in $\mathbf{R}^3$ the sum of the four unit normals to a regular tetrahedron.

3.9. (a) Calculate in $\mathbf{R}^3$ the Minkowski sum of the six segments $[(0, 0, 0), (1, 0, 0)]$, $[(0, 0, 0), (0, 1, 0)]$, $[(0, 0, 0), (0, 0, 1)]$, $[(0, 0, 0), (0, 1, 1)]$, $[(0, 0, 0), (1, 0, 1)]$, and $[(0, 0, 0), (1, 1, 0)]$. How many vertices, edges, and faces are there?

(b) Describe the Minkowski sum of the six edges of a regular tetrahedron in $\mathbf{R}^3$.

3.10. Prove that a triangle cannot be the Minkowski sum of a finite number of segments.

3.11. A heptagon in $\mathbf{R}^2$ can never be the sum of a finite number of segments.

3.12. A tetrahedron cannot be the sum of a finite number of segments.

3.13. The vector sum of two triangles in $\mathbf{R}^2$ is a polygon. How many sides can it have? How many sides can the sum of a triangle and a square have?

3.14. Give an example of a quadrilateral in $\mathbf{R}^2$ that is not the sum of two segments.

4. Closure operators on subsets of a semigroup

4.1. Combining erosions and dilations

We have defined in Chapter 3 dilations and erosions. We shall now combine them.

It is clear that $(A + B) + C = A + (B + C)$, so the composition of two dilations is a dilation: $\delta_C \circ \delta_B = \delta_{B+C}$. Similarly, $(A \ominus B) \ominus C = A \ominus (B + C)$, so the composition of two erosions is an erosion: $\varepsilon_C \circ \varepsilon_B = \varepsilon_{B+C}$. (In the non-commutative case the conclusions are the same.) But what about the composition of a dilation and an erosion? Let us consider $\gamma = \varepsilon_C \circ \delta_B$, defined by $\gamma(A) = (A + B) \ominus C$, where B and C are fixed subsets of an abelian group. This operation is always increasing, and it is extensive if and only if $C \subset B$, which in turn is equivalent to $\gamma(\{0\}) \supset \{0\}$. In this section we shall study this operator when $C = B$ (this is the only really interesting case) and see that it is idempotent then (cf. exercises 4.2 and 4.3). We shall also see that, in typical cases, it is neither a dilation nor an erosion.

Given two subsets A and B of a semigroup G we define

$$\gamma_B(A) = \varepsilon_B(\delta_B(A)) = A^B = \{x \in G; xB \subset AB\}, \qquad A \in \mathscr{P}(G); \tag{4.1.1}$$

in the commutative case $\gamma_B(A) = \{x \in G; x + B \subset A + B\}$. This means that we perform on A first a dilation by B, then an erosion by the same set. We shall see that this is a closure operator in the sense of Chapter 2. So we call $\gamma_B(A) = A^B$ the *B-closure* of A, and a set A is called *B-closed* if $A^B = A$.[11]

Analogously we define

$$(4.1.2); \qquad \eta_B(A) = \delta_B(\varepsilon_B(A)) = A_B = \bigcup_{x \in G}(xB; xB \subset A), \qquad A \in \mathscr{P}(G),$$

in the commutative case $\eta_B(A) = \bigcup(x + B; x + B \subset A)$. Here we perform first an erosion by B and then a dilation by B. We call $\eta_B(A) = A_B$ the *B-opening* of A, and a set A is called *B-open* if $A_B = A$.[12]

Theorem 4.1.1. *For any subset B of an arbitrary semigroup G the mapping $\gamma_B = \varepsilon_B \circ \delta_B$, $\mathscr{P}(G) \ni A \mapsto A^B \in \mathscr{P}(G)$, is a closure operator, and the mapping $\eta_B = \delta_B \circ \varepsilon_B$, $A \mapsto A_B$, is an opening.*

The theorem will follow on combining Proposition 4.1.2 and Corollary 4.1.4 below.

Proposition 4.1.2. *For any subset B of a semigroup G the mappings $\gamma_B = \varepsilon_B \circ \delta_B$, $A \mapsto A^B$, and $\eta_B = \delta_B \circ \varepsilon_B$, $A \mapsto A_B$, are increasing; the first is extensive and the second is antiextensive.*

Proof. It is obvious that the mappings are increasing. Formula (4.1.1) shows that A^B contains A; similarly we see from (4.1.2) that A_B is contained in A.

Corollary 4.1.3. *For any subset B of a semigroup G we have*

$$\delta_B \circ \varepsilon_B \circ \delta_B = \delta_B \text{ and } \varepsilon_B \circ \delta_B \circ \varepsilon_B = \varepsilon_B.$$

Proof. Let us write composition as juxtaposition for brevity and omit the subscript B. Then $\delta\varepsilon\delta \leqslant \delta$ since $\delta\varepsilon \leqslant \mathsf{Id}$. But we also have $\delta\varepsilon\delta \geqslant \delta$ since $\varepsilon\delta \geqslant \mathsf{Id}$ and δ is increasing.

Similarly, $\varepsilon\delta\varepsilon \leqslant \varepsilon$ since $\delta\varepsilon \leqslant \mathsf{Id}$ and ε is increasing. But we also get $\varepsilon\delta\varepsilon \geqslant \varepsilon$ since $\varepsilon\delta \geqslant \mathsf{Id}$.

Corollary 4.1.4. *For any subset B of a semigroup G the mappings $\gamma_B = \varepsilon_B \circ \delta_B$ and $\eta_B = \delta_B \circ \varepsilon_B$ are idempotent.*

[11] Matheron (1975) and Serra (1982, 2001) use the notation A^B; Gonzalez & Woods (1993:524) write $A \bullet B$.

[12] Matheron (1975) and Serra (1982, 2001) use the notation A_B, Gonzalez & Woods (1993:524) use $A \circ B$. Note, however, that $\circ$ is already used to denote an operation in mathematics, viz. composition of functions or relations.

Proof. We have $\gamma\gamma = (\varepsilon\delta)(\varepsilon\delta) = (\varepsilon\delta\varepsilon)\delta = \varepsilon\delta = \gamma$. Similarly $\eta\eta = (\delta\varepsilon)(\delta\varepsilon) = (\delta\varepsilon\delta)\varepsilon = \delta\varepsilon = \eta$.

Corollary 4.1.5. *For any subsets A, B of a semigroup G, $\delta_B(A) = AB$ is B-open and $\varepsilon_B(A)$ is B-closed.*

Proof. We have $\eta\delta = \delta\varepsilon\delta = \delta$, so $\eta_B(\delta_B(A)) = \delta_B(A)$, showing that $\delta_B(A)$ is B-open. Similarly $\gamma\varepsilon = \varepsilon\delta\varepsilon = \varepsilon$ and $\gamma_B(\varepsilon_B(A)) = \varepsilon_B(A)$; $\varepsilon_B(A)$ is B-closed.

For any fixed B, the family of all sets $\varepsilon_B(A)$ is equal to the family of all B-closed sets. This is the same as the family of all sets $\gamma_B(A)$. Suppose we put on spectacles blocking out everything except the B-closed sets. The whole world becomes B-closed: our B-spectacles do not permit us to see anything else. Then we can only see sets of the form $A \ominus B$. However, A^B is also B-closed and usually a much better approximation of A than $A \ominus B$.

If G is a group, ε_B is the mapping dual to $\delta_{\check{B}}$, i.e., $\varepsilon_B = (\delta_{\check{B}})^{\mathbf{d}} = \complement \circ \delta_{\check{B}} \circ \complement$, so we have

$$\eta_B = \delta_B \circ \varepsilon_B = \delta_B \circ \complement \circ \delta_{\check{B}} \circ \complement = \complement \circ \varepsilon_{\check{B}} \circ \delta_{\check{B}} \circ \complement = \complement\gamma_{\check{B}}\complement = (\gamma_{\check{B}})^{\mathbf{d}},$$

so that η_B and $\gamma_{\check{B}}$ are dual to each other. This implies that $\complement\,(A_B) = \big(\complement A\big)^{\check{B}}$ and shows that a set is B-open if and only if its complement is $\check{B}$-closed.

4.2. Characterizing closure operators which commute with translations

In an abelian group the mappings that commute with translations are of special significance. We can characterize closure operators which commute with translations in terms of Minkowski addition as follows.

Theorem 4.2.1. *Let G be an abelian group and $\mathscr{A}$ a subfamily of $\mathscr{P}(G)$ such that $\mathscr{A}$ contains all singleton sets $\{x\}$, $x \in G$, and such that $A + B \in \mathscr{A}$ for all $A, B \in \mathscr{A}$. Let $f\colon \mathscr{A} \to \mathscr{P}(G)$ be a closure operator in the sense of Definition 2.2.2. Then the following conditions are equivalent.*
(A) *f commutes with all translations: $f \circ T_x = T_x \circ f$ for all $x \in G$;*
(B) *$f(A + x) \supset f(A) + x$ for all $A \in \mathscr{A}$ and all $x \in G$;*
(C) *$f(A + B) \supset f(A) + B$ for all $A, B \in \mathscr{A}$ (thus $f \circ \delta_B \supset \delta_B \circ f$ for all $B \in \mathscr{A}$);*
(D) *$f(A + B) \supset f(A) + f(B)$ for all $A, B \in \mathscr{A}$.*

Proof. That (A) and (B) are equivalent is easily proved.

Assume that (B) holds. Then $f(A+B) \supset f(A+y) \supset f(A)+y$ for all $y \in B$. Now (C) follows on taking the union over all y.

If (C) holds we know that $f(A + B) \supset A + f(B)$, which implies that $f(A + B) \supset f(A+f(B))$ by (2.2.5). Applying (C) a second time we see that the latter set contains $f(A) + f(B)$; hence (D) holds.

Finally, if (D) holds, then $f(A + x) \supset f(A) + f(\{x\}) \supset f(A) + x$. Thus (B) holds.

In the case of a vector space and a mapping which commutes with homotheties, thus $f(\lambda A) = \lambda f(A)$ for nonnegative λ, it follows from (D) that

$$f(\lambda A + \mu B) \supset \lambda f(A) + \mu f(B) \text{ for all } \lambda, \mu \geqslant 0 \text{ and all } A, B \in \mathscr{A},$$

which is a property analogous to concavity.[13]

4.3. Matheron's structural theorems

We are now acquainted with dilations, erosions, closings, and openings. They are examples of morphological mappings, but we have so far no idea how general they are. Are they just some special mappings that we have come across? There are obviously many more mappings $\mathscr{P}(G) \to \mathscr{P}(G)$, even if we restrict ourselves to mappings which commute with translations, which is reasonable to do. More precisely, we ask the following two questions.

How special are the dilations δ_B and erosions ε_B that we have studied so far in the family of all increasing mappings which commute with translations?

How special are the closings γ_B and the openings η_B we have constructed in the family of all closings and openings which commute with translations?

Georges Matheron proved two results which gave very neat answers to these two questions.

Matheron's first structural theorem (1975: Proposition 8-1-3) describes the structure of increasing mappings in $\mathbf{R}^n$ which commute with translations in terms of dilations and erosions. Actually the result holds in every abelian group. It is important because it underlines the fact that the mappings we already know are the building blocks of a much more general class of mappings.

For any mapping $f\colon \mathscr{P}(G_1) \to \mathscr{P}(G_2)$ of the family of all subsets of an abelian group G_1 into those of another, G_2, we define its *kernel* as

$$\ker f = \{A \in \mathscr{P}(G_1); 0 \in f(A)\}.$$

The notion is due to Matheron (1975:217). A mapping which commutes with all translations is completely determined by its kernel.

We mention some examples of kernels.

1. The kernel of the identity $\mathsf{Id}\colon \mathscr{P}(G) \to \mathscr{P}(G)$ is $\ker \mathsf{Id} = \{A \in \mathscr{P}(G); 0 \in A\}$.
2. The kernel of the mapping $f^{\mathbf{d}}$ dual to f is $\ker f^{\mathbf{d}} = \{A; 0 \notin f\left(\complement A\right)\}$.
3. The kernel of a translation T_b, $T_b(A) = A + b$, is $\ker T_b = \{A; -b \in A\}$.
4. A dilation $\delta_B\colon A \mapsto A + B$ has kernel $\ker \delta_B = \{A; A \cap \check{B} \neq \emptyset\}$.
5. The kernel of an erosion $\varepsilon_B\colon A \mapsto A \ominus B$ is $\ker \varepsilon_B = \{A; A \supset B\}$.
6. The kernel of a closure operator $\gamma_B = \varepsilon_B \circ \delta_B$ is
$$\ker \gamma_B = \{A; B \subset A + B\} = \bigcap_{y \in B} \ker \delta_{B-y}.$$
7. An opening $\eta_B = \delta_B \circ \varepsilon_B\colon A \mapsto A_B$ has kernel
$$\ker \eta_B = \{A; \exists y \in B \text{ such that } A \supset B - y\} = \bigcup_{y \in B} \ker \varepsilon_{B-y}.$$

[13] This property was introduced and used to study closure operators appearing in the theory of partial differential equations by Kiselman (1969); however, I did not know at the time that it is implied by (A).

Theorem 4.3.1 (Matheron's first structural theorem). *Let G be an abelian group and $f\colon \mathscr{P}(G) \to \mathscr{P}(G)$ an increasing mapping which commutes with translations. Then f is a union of erosions as well as an intersection of dilations:*

$$f(A) = \bigcup_{B \in \ker f} \varepsilon_B(A) = \bigcup_{B \in \ker f} (A \ominus B) = \bigcap_{B \in \ker f^{\mathbf{d}}} \delta_{\check{B}}(A) = \bigcap_{B \in \ker f^{\mathbf{d}}} (A - B), \tag{4.3.1}$$

for all $A \in \mathscr{P}(G)$, where $f^{\mathbf{d}}(A) = \complement f(\complement A)$ is the mapping dual to f.

Proof. If $B \in \ker f$, then $0 \in f(B)$ and, since f commutes with all translations, $x \in f(x + B)$ for all $x \in G$. Therefore, given any subset A of G, $x \in \varepsilon_B(A)$ implies that $x + B \subset A$ and $x \in f(x + B) \subset f(A)$ (f is increasing). Since x is arbitrary in $\varepsilon_B(A)$, we have proved that $\varepsilon_B(A) \subset f(A)$ for all A and all $B \in \ker f$. Letting B vary in $\ker f$ we see that the union of all the $\varepsilon_B(A)$ is contained in $f(A)$.

To prove the inclusion in the other direction, take an arbitrary element x of $f(A)$. Then $0 = x - x \in f(A) - x = f(A - x)$, i.e., $A - x \in \ker f$. If we now define B as $A - x$, then $x \in \varepsilon_B(A)$. This means that there exists a $B \in \ker f$ such that $x \in \varepsilon_B(A)$, and we have proved that $f(A)$ is contained in the union of all the $\varepsilon_B(A)$.

The second representation follows on applying the first to $f^{\mathbf{d}}$.

Matheron's second structural theorem requires a preliminary study of extensions of increasing mappings which we shall now undertake.

Let X be an arbitrary set and $\mathscr{A}$ a subset of $\mathscr{P}(X)$. For any mapping $f\colon \mathscr{A} \to \mathscr{P}(X)$ we then define

$$f_\diamond(B) = \bigcup_{\substack{A \in \mathscr{A} \\ A \subset B}} f(A) \text{ and } f^\diamond(B) = \bigcap_{\substack{A \in \mathscr{A} \\ A \supset B}} f(A), \qquad B \in \mathscr{P}(X). \tag{4.3.2}$$

Lemma 4.3.2. *Let $\mathscr{A}$ be any subset of $\mathscr{P}(X)$, where X is an arbitrary set. The mappings $f_\diamond$ and $f^\diamond$ are dual to each other in the sense that $(f_\diamond)^{\mathbf{d}} = (f^{\mathbf{d}})^\diamond$ and $(f^\diamond)^{\mathbf{d}} = (f^{\mathbf{d}})_\diamond$. If $f\colon \mathscr{A} \to \mathscr{P}(X)$ is an increasing mapping, then $f_\diamond$ is the smallest increasing extension of f to all of $\mathscr{P}(X)$, and $f^\diamond$ is the largest increasing extension of f to all of $\mathscr{P}(X)$.*

Proof. To prove the duality result is straightforward. If f is increasing, it follows that $f_\diamond$ is actually an extension of f, i.e., that $f_\diamond\big|_{\mathscr{A}} = f$. If g is an arbitrary increasing extension, we see that we must have $g(B) \supset f_\diamond(B)$ for all B, proving that $f_\diamond$ is the smallest increasing extension. The result for $f^\diamond$ follows by duality.

Proposition 4.3.3. *Let X be any set, $\mathscr{A}$ a subset of $\mathscr{P}(X)$, and let $g\colon \mathscr{A} \to \mathscr{P}(X)$ be an opening in the sense of Definition 2.2.4. Assume that $g(A) \in \mathscr{A}$ for every $A \in \mathscr{A}$. Then $g_\diamond\colon \mathscr{P}(X) \to \mathscr{P}(X)$ is an opening.*

Proof. That $g_\diamond$ is increasing and antiextensive is obvious. It remains to be proved that $g_\diamond$ is idempotent. We know that $g_\diamond$ is an extension of g, which implies that $g_\diamond(g(A)) = g(g(A)) = g(A)$ for all $A \in \mathscr{A}$, since by hypothesis $g(A) \in \mathscr{A}$. Hence $g_\diamond(g_\diamond(B)) \supset g_\diamond(g(A)) = g(A)$ if $A \in \mathscr{A}$ and $A \subset B$. On taking the union over all

$A \in \mathscr{A}$ contained in B we get

$$g_\diamond(g_\diamond(B)) \supset \bigcup_{\substack{A \in \mathscr{A} \\ A \subset B}} g(A) = g_\diamond(B) \supset g_\diamond(g_\diamond(B)),$$

showing that $g_\diamond$ is idempotent.

Remark. The hypothesis that g maps $\mathscr{A}$ into itself cannot be dispensed with. Indeed, take $X = [0, +\infty[$, $A_j = [0, j]$ for $j = 0, 1, 2, 3$, and define $g(A_j) = A_{j-1} = [0, j-1]$, $j = 1, 3$, and $\mathscr{A} = \{A_1, A_3\}$. Then $g \colon \mathscr{A} \to \mathscr{P}(X)$ is an opening in the sense of Definition 2.2.4; in particular (2.2.11) holds: $g(A_j) \subset A_k$ implies $g(A_j) \subset g(A_k)$ for $j, k = 1, 3$. But $g_\diamond$ is not idempotent: $A_0 = g_\diamond(g_\diamond(A_3)) \neq g_\diamond(A_3) = A_2$.

For any mapping $f \colon \mathscr{A} \to \mathscr{P}(G)$ we shall say that a set A is *f-invariant* if $f(A) = A$, and we shall denote by $\mathsf{Inv}_f = \{A \in \mathscr{A}; f(A) = A\}$ the set of all f-invariant sets, the *invariance set of f*. If f is a closing, this is the set of all f-closed elements; if f is an opening, it is the set of all f-open elements.

Proposition 4.3.4. *If $g \colon \mathscr{A} \to \mathscr{P}(X)$ is an opening, then Inv_g is closed under the formation of unions (in particular $\varnothing \in \mathsf{Inv}_g$), and $g = \left(\mathsf{Id}_{\mathsf{Inv}_g}\right)_\diamond\big|_{\mathscr{A}}$, i.e., g is the smallest increasing extension to $\mathscr{P}(X)$ of the identity on Inv_g. Conversely, let a class $\mathscr{C} \subset \mathscr{P}(X)$ be given. Then $(\mathsf{Id}_{\mathscr{C}})_\diamond$, the smallest increasing extension of the identity on $\mathscr{C}$, is an opening h, and Inv_h is the class closed under the formation of unions generated by $\mathscr{C}$.*

Proof. If $A_j \in \mathscr{A}$, $j \in J$, are g-invariant, then $g(\bigcup A_j) \supset g(A_k) = A_k$ for all k, so $g(\bigcup A_j) \supset \bigcup A_k$. Since on the other hand we always have $g(\bigcup A_j) \subset \bigcup A_j$, it follows that $\bigcup A_j \in \mathsf{Inv}_g$.

That g is the smallest increasing extension of the identity on Inv_g follows from

$$\left(\mathsf{Id}_{\mathsf{Inv}_g}\right)_\diamond(B) = \bigcup_{\substack{A \in \mathscr{A} \\ A \subset B}} A = B \text{ if } B \in \mathscr{A}.$$

Since the inclusion mapping $\mathscr{C} \to \mathscr{P}(X)$ is an opening satisfying the condition in Proposition 4.3.3, it follows that $(\mathsf{Id}_{\mathscr{C}})_\diamond$ is an opening. To prove the last assertion in the statement of the proposition, we note that by the definition of $(\mathsf{Id}_{\mathscr{C}})_\diamond$, $h(B) = B$ implies that $B = h(B) = \bigcup(A; A \in \mathscr{C}, A \subset B)$, which means that B belongs to the set closed under union formation generated by $\mathscr{C}$.

In case $X = G$ is an abelian group, let us say that a subset $\mathscr{B}$ of Inv_g is a *basis for* Inv_g if Inv_g is the class closed under formation of unions and translations generated by $\mathscr{B}$. This means that Inv_g is the family of all sets of the form $\bigcup_{(B,x)\in M}(B + x)$ for some subset M of $\mathscr{B} \times G$.

We are now ready to state and prove Matheron's second structural theorem (1975: Proposition 7-1-3), which characterizes general closings and openings in terms of the elementary closings and openings γ_B and η_B by a given subset B. His result, which shows that the latter are not so special as one could imagine, is as follows.

Theorem 4.3.5 (Matheron's second structural theorem). *Let G be an abelian group and $g\colon \mathscr{P}(G) \to \mathscr{P}(G)$ a mapping which commutes with translations. Then g is an opening if and only if it admits a representation*

$$g(A) = \bigcup_{B\in\mathscr{B}} \eta_B(A) = \bigcup_{B\in\mathscr{B}} A_B$$

for some class $\mathscr{B} \subset \mathscr{P}(G)$. If this is so, the class of sets invariant under g is the family of all elements of $\mathscr{B}$ and all translates and unions of these. If g is an opening, the dual mapping $g^{\mathbf{d}}$ is a closing and has the representation

$$g^{\mathbf{d}}(A) = \bigcap_{B\in\mathscr{B}} \gamma_B(A) = \bigcap_{B\in\mathscr{B}} A^B.$$

Proof. First assume that g is an opening. The smallest extension of the identity on Inv_g is given by

$$g(A) = \left(\mathsf{Id}_{\mathsf{Inv}_g}\right)_\diamond(A) = \bigcup_{\substack{C\in\mathsf{Inv}_g\\ C\subset A}} C = \bigcup_{B\in\mathscr{B}}\bigcup_{x\in G}(B+x; B+x\subset A), \tag{4.3.3}$$

if $\mathscr{B}$ is a basis for Inv_g as defined before the statement of the theorem. In fact, given a set $C \in \mathsf{Inv}_g$ contained in A, there is a subset M of $\mathscr{B}\times G$ such that

$$C = \bigcup_{(B,x)\in M} (B+x; B+x\subset A) \subset \bigcup_{(B,x)\in\mathscr{B}\times G} (B+x; B+x\subset A) = C',$$

but since C' is also admissible, the union will not be changed if we add it, and (4.3.3) follows. Let us now note that

$$A_B = (A \ominus B) + B = \bigcup_{x\in G}(B+x; B+x\subset A).$$

Therefore the last expression in (4.3.4) is equal to $\bigcup_{B\in\mathscr{B}} A_B$.

Conversely, if g_j, $j\in J$, are openings, so is g defined by $g(A) = \bigcup g_j(A)$. Indeed, g is certainly increasing and antiextensive, and idempotency follows from

$$g(g(A)) = \bigcup_{j\in J} g_j\left(\bigcup_{k\in J} g_k(A)\right) \supset \bigcup_{j\in J} g_j(g_j(A)) = \bigcup_{j\in J} g_j(A) = g(A).$$

4.4. Exercises

4.1. Find examples (for instance in $\mathbf{Z}^2$) that show that we do not always have $\gamma_B(A_1\cup A_2) = \gamma_B(A_1)\cup\gamma_B(A_2)$ or $\gamma_B(A_1\cap A_2) = \gamma_B(A_1)\cap\gamma_B(A_2)$; in other words, $\gamma_B = \varepsilon_B\circ\delta_B$ is in general neither a dilation nor an erosion.

4.2. Define in an abelian group G a mapping $\eta = \delta_C\circ\varepsilon_B$, thus $\eta(A) = (A\ominus B)+C$, $A\in\mathscr{P}(G)$, where B and C are fixed subsets of G. Prove that η is antiextensive if and only if $C\subset B$, which in turn is equivalent to $\eta(G\smallsetminus\{0\})\subset G\smallsetminus\{0\}$. Prove that if B and C are nonempty bounded subsets of $\mathbf{Z}^n$ or $\mathbf{R}^n$ with $C \subset B$, then η is idempotent only if $B = C$, thus only if $\eta(A) = A_B$. *Hint:* Calculate $\eta(B)$ and $\eta(\eta(B))$.

4.3. Study the dual mapping $\gamma = \varepsilon_{\check{C}} \circ \delta_{\check{B}}$.

5. Distance transforms

5.1. Definition and basic properties of distance transforms

Distance transforms of digital images are a useful tool in image analysis. The distance transform of a set (or shape, or image) is a function on the image carrier. Outside the set, the value of the distance transform at a certain pixel is defined to be the distance from that pixel to the set. Inside the set, it is often defined as the distance to the complement, but we shall find it convenient to define it instead as minus the distance to the complement, for a very simple reason: the distance transform of a convex set in $\mathbf{R}^n$ is a convex function with this definition.

The distances can be measured in different ways, e.g., by approximating the Euclidean distance in the two-dimensional image, the Euclidean distance between two pixels $x = (x_1, x_2)$ and $y = (y_1, y_2)$ being

$$d^2(x,y) = \|x-y\|_2 = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}.$$

Other distances that have been used are the *city-block distance* or l^1*-distance*

$$d^1(x,y) = \|x-y\|_1 = |x_1-y_1| + |x_2-y_2|$$

and the *chessboard distance* or l^∞*-distance*

$$d^\infty(x,y) = \|x-y\|_\infty = \max(|x_1-y_1|, |x_2-y_2|).$$

We shall define many more distances on $\mathbf{Z}^n$ in section 5.3.

Let X be any nonempty set. Let us agree to call a function $d\colon X \times X \to \mathbf{R}$ a *distance* if d is *positive definite*:

$$d(x,y) \geqslant 0 \text{ with equality precisely when } x = y, \qquad x, y \in X, \tag{5.1.1}$$

and *symmetric*:

$$d(x,y) = d(y,x) \text{ for all } x, y \in X. \tag{5.1.2}$$

A distance will be called a *metric* if in addition it satisfies the *triangle inequality*:

$$d(x,z) \leqslant d(x,y) + d(y,z) \text{ for all } x, y, z \in X. \tag{5.1.3}$$

Every nonempty set can be equipped with a metric, viz. the discrete metric d_0 defined as

$$d_0(x,x) = 0, \qquad d_0(x,y) = 1 \text{ if } x \neq y. \tag{5.1.4}$$

The set X will usually be the image plane $\mathbf{Z}^2$ consisting of all points in the plane with integer coordinates (the addresses of the pixels), or more generally the image space $\mathbf{Z}^n$.

Whenever X is an abelian group it is of particular interest to use *translation-invariant* distances, i.e., those which satisfy

$$d(x-a, y-a) = d(x,y) \text{ for all } a, x, y \in X. \tag{5.1.5}$$

A *metric space* is simply a set provided with a metric. In any metric space we define balls as follows: the *closed ball of center c and radius r* (or the *non-strict ball*) is the set of all points x satisfying $d(x,c) \leqslant r$ and will be denoted by $B_\leqslant(c,r)$; the *open ball of center c and radius r* (or the *strict ball*) is the set of all points satisfying the strict inequality $d(x,c) < r$; it will be denoted by $B_<(c,r)$. In the Euclidean case these notions are well known, but in general we must be a bit careful: the closure of $B_<(c,r)$ with respect to the topology defined by d is not necessarily equal to $B_\leqslant(c,r)$, and the interior of $B_\leqslant(c,r)$ is not necessarily equal to $B_<(c,r)$ (see section 5.5). Also note that if two balls $B_<(c_1,r_1)$ and $B_<(c_2,r_2)$ with $r_1, r_2 > 0$ are disjoint, then we can only conclude that $\max(r_1,r_2) \leqslant d(c_1,c_2)$, whereas in a normed space a stronger inequality, $\max(r_1,r_2) \leqslant r_1 + r_2 \leqslant \|c_1 - c_2\|$, holds.

Every metric defines a topology: a set is declared to be open if and only if it is a union of open balls. However, we shall often use another topology on X than that defined by d.

We note that in any abelian group with a translation-invariant metric we have the relations

$$B_<(c_1,r_1) + B_<(c_2,r_2) \subset B_\leqslant(c_1,r_1) + B_<(c_2,r_2) \subset B_<(c_3,r_3);$$

$$B_\leqslant(c_1,r_1) + B_\leqslant(c_2,r_2) \subset B_\leqslant(c_3,r_3),$$

where $c_1 + c_2 = c_3$, $r_1 + r_2 = r_3$. In a vector space, with d defined by $d(x,y) = \|x - y\|$ using some norm $\|\cdot\|$, the inclusions here are actually equalities if $r_1, r_2 > 0$.

Definition 5.1.1. *In a metric space X we define the* distance transform DT_A *of a subset A of X by*

$$\mathsf{DT}_A(x) = \begin{cases} -\inf\limits_{y \notin A} d(x,y), & x \in A; \\ \inf\limits_{y \in A} d(x,y), & x \in X \smallsetminus A. \end{cases} \tag{5.1.6}$$

Lemma 5.1.2. *The distance transform satisfies*

$$\mathsf{DT}_A(x) = \begin{cases} -\sup\big(r; B_<(x,r) \subset A\big), & x \in A; \\ \sup\big(r; B_<(x,r) \subset \complement A\big), & x \in X \smallsetminus A. \end{cases} \tag{5.1.7}$$

Proof. If $x \notin A$, $y \in A$, and $B_<(x,r) \subset \complement A$, then $r \leqslant d(x,y)$. This shows that the supremum of all such r cannot exceed the infimum of all $d(x,y)$ with $y \in A$. On the other hand, if r_1 is larger than the supremum of all r such that $B_<(x,r)$ does not meet A, then there exists a point $z \in A$ such that $d(x,z) < r_1$, which shows that the infimum of $d(x,z)$ over all such z is less than r_1, hence less than or equal to $\sup r$. This proves the result when $x \notin A$. The case when $x \in A$ is similar.

Thus when x does not belong to A, the distance transform $\mathsf{DT}_A(x)$ is what we naturally understand by the distance from x to A, but we complement this idea by defining also the distance transform inside A. It is then natural to take it negative there, so that A is approximately the set where the transform is negative. As already remarked it is convenient that, in a normed space, the distance transform of a convex set is a convex function; see Corollary 5.6.3.

Note the symmetry: $\mathsf{DT}_{X\smallsetminus A} = -\mathsf{DT}_A$. The distance transformation $A \mapsto \mathsf{DT}_A$ is decreasing in the sense that $\mathsf{DT}_A(x) \geqslant \mathsf{DT}_B(x)$ for all $x \in X$ if $A \subset B$.

In the two extreme cases $A = \emptyset$ and $A = X$ we have $\mathsf{DT}_\emptyset = +\infty$ and $\mathsf{DT}_X = -\infty$. In all other cases DT_A is real-valued, $\mathsf{DT}_A \colon X \to \mathbf{R}$.

Every real-valued function can be written as the difference between two nonnegative functions: $f = f^+ - f^-$, where $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$. In particular, $\mathsf{DT}_A = (\mathsf{DT}_A)^+ - (\mathsf{DT}_A)^-$. The function $(\mathsf{DT}_A)^-$ is sometimes called the *quench function* of A.[14]

Proposition 5.1.3. *If A is a subset of a metric space X other than $\emptyset$ and X, then $(\mathsf{DT}_A)^+$ and $(\mathsf{DT}_A)^-$ are Lipschitz continuous with Lipschitz constant* 1 *with respect to d:*

$$|(\mathsf{DT}_A)^+(x) - (\mathsf{DT}_A)^+(y)| \leqslant d(x,y), \qquad x, y \in X, \tag{5.1.8}$$

and similarly for $(\mathsf{DT}_A)^-$. (In particular the restrictions $\mathsf{DT}_A\big|_A$ and $\mathsf{DT}_A\big|_{\complement A}$ are Lipschitz continuous with Lipschitz constant 1*.) As a consequence, DT_A is Lipschitz continuous with Lipschitz constant* 2*. If X is a vector space with distance $d(x-y) = \|x-y\|$ defined by a norm, the Lipschitz constant is* 1.

Let us say for brevity that a function is *Lip*-1 if it is Lipschitz continuous with Lipschitz constant 1, i.e., when it satisfies (5.1.8).

Proof. The restriction to A of DT_A is the supremum of a family of Lip-1 functions $x \mapsto d(x,y)$; it is easy to prove that this operation preserves the Lipschitz constant. It follows that $|(\mathsf{DT}_A)^-(a) - (\mathsf{DT}_A)^-(b)| \leqslant d(a,b)$ if $a, b \in A$. If $a, b \notin A$ the function takes the value zero at both points.

Now take $a \in A$ and $b \in X \smallsetminus A$ and define $r = -\mathsf{DT}_A(a) \geqslant 0$ and $s = \mathsf{DT}_A(b) \geqslant 0$. Then the open ball $B_<(a,r)$ is contained in A, and the open ball $B_<(b,s)$ is contained in $\complement A$, so that $r, s \leqslant d(a,b)$. The two balls are disjoint. In general this only implies that $\max(r,s) \leqslant d(a,b)$, but in a normed vector space case the stronger inequality $s + r \leqslant d(a,b)$ follows, thus that $0 \leqslant \mathsf{DT}_A(b) - \mathsf{DT}_A(a) = s + r \leqslant d(a,b)$; proving that the Lipschitz constant is 1 in this case.

Returning to the general case, we note that, when $a \in A$ and $b \notin A$, we have $0 = -(\mathsf{DT}_A)^-(b) \leqslant \mathsf{DT}_A(a)^- - \mathsf{DT}_A(b)^- = r \leqslant d(a,b)$ and the Lipschitz continuity of $(\mathsf{DT}_A)^-$ is established. Passing to the complement, we obtain the result for $(\mathsf{DT}_A)^+ = (\mathsf{DT}_{X\smallsetminus A})^-$.

The difference $\mathsf{DT}_A = (\mathsf{DT}_A)^+ - (\mathsf{DT}_A)^-$ thus has Lipschitz constant at most 2.

The Lipschitz constant 2 in the proposition cannot be improved as can be seen from the simplest of examples: $X = \mathbf{Z}$ with the usual metric and $A = \{0\}$. However, in the distance transform there is a jump 2 only when we go from a point in A to a point in

[14] Serra (1982:377), who attributes the term to L. Calabi. However, it seems not to be widely used.

$X \smallsetminus A$. This indicates that it might be possible to adjust the distance transform in A by an additive constant so that the modified function is Lip-1. The following results makes this idea explicit.

Theorem 5.1.4. *For a subset A in a metric space X we define two quantities*

$$\alpha(A) = \sup\bigl(r+s-d(a,b)\bigr) \text{ and } \beta(A) = \inf\bigl(r+s+d(a,b)\bigr),$$

where the supremum and infimum are taken over all points $a \in A$, $b \notin A$, and all balls $B_<(a,r)$, $B_<(b,s)$ that are disjoint and maximal with this property. We also define a modified distance transform as

$$F_{A,\omega}(x) = \begin{cases} \mathsf{DT}_A(x) + \omega, & x \in A; \\ \mathsf{DT}_A(x), & x \in X \smallsetminus A. \end{cases} \tag{5.1.9}$$

Then if A and $X \smallsetminus A$ are nonempty and $\alpha(A) \leqslant \beta(A)$, there exists a real number ω such that $F_{A,\omega}$ is Lip-1; we can take ω as any number satisfying $\alpha(A) \leqslant \omega \leqslant \beta(A)$. Conversely, if the function $F_{A,\omega}$ is Lip-1 for some ω, then $\alpha(A) \leqslant \beta(A)$.

Proof. We have to prove that

$$|F_{A,\omega}(b) - F_{A,\omega}(a)| \leqslant d(a,b), \qquad a,b \in X. \tag{5.1.10}$$

The proof of Proposition 5.1.3 shows that the inequality holds if $a,b \in A$ as well as if $a,b \notin A$. The only question left is when $a \in A$ and $b \notin A$. Then we have

$$|F_{A,\omega}(b) - F_{A,\omega}(a)| = |s - (-r+\omega)| = |r+s-\omega|, \tag{5.1.11}$$

where r and s are defined as in the proof of Proposition 5.1.3. If $\alpha(A) \leqslant \omega \leqslant \beta(a)$ we know that for all a,b,r,s under consideration we have $r+s-d(a,b) \leqslant \alpha(A)$ and $r+s+d(a,b) \geqslant \beta(A)$. With any ω betwen $\alpha(A)$ and $\beta(A)$ we thus get

$$r+s-d(a,b) \leqslant \omega \leqslant r+s+d(a,b),$$

which may be written as $|r+s-\omega| \leqslant d(a,b)$. With this inequality, (5.1.11) implies (5.1.10), which means that the modified distance transform is Lip-1.

Conversely, if (5.1.10) holds for all a,b,r,s that we consider, then $|r+s-\omega| \leqslant d(a,b)$ must be true, which leads to the inequality $\alpha(A) \leqslant \beta(A)$.

In a vector space we have $\alpha(A) = \beta(A) = 0$ for every nonempty $A \neq X$. Hence $\omega = 0$ is the only choice; this is already clear from Proposition 5.1.3.

In $\mathbf{Z}^n$ with the l^∞ metric, we have $\alpha(A) = 1$, $\beta(A) = 3$ for all $A \neq \varnothing, \mathbf{Z}^n$, so that any $\omega \in [1,3]$ will do.

In general the condition $\alpha(A) \leqslant \beta(A)$ for all A is a strong regularity condition on the metric space X. For instance, if there exist points a_0 and b_0 in X such that $r+s > d(a_0,b_0)$, then $\alpha(A) > 0$ for any set A such that $a_0 \in A$ and $b_0 \notin A$. The condition $\alpha(A) \leqslant \beta(A)$ implies that points in A and $\complement A$ must not come too close: $\alpha(A) \leqslant \beta(A) \leqslant r+s+d(a,b) \leqslant 3d(a,b)$, so that $d(a,b) \geqslant \frac{1}{3}\alpha(A) > 0$ for all sets A such that $a_0, a \in A$ and $b_0, b \in \complement A$.

Proposition 5.1.5. *Let G be an abelian group with a translation-invariant metric $d(x,y) = f(x-y)$, and let A be an arbitrary subset of G. We denote by i_A the indicator function of A. Then*

$$(\mathsf{DT}_A)^+ = \max(\mathsf{DT}_A, 0) = i_A \mathbin{\square} f \text{ and } (\mathsf{DT}_A)^- = \max(-\mathsf{DT}_A, 0) = i_{\complement A} \mathbin{\square} f,$$

and, taking the difference between the two,

$$\mathsf{DT}_A = (\mathsf{DT}_A)^+ - (\mathsf{DT}_A)^- = (i_A \mathbin{\square} f) - (i_{\complement A} \mathbin{\square} f).$$

Proof. We see that in $X \smallsetminus A$ we have $\mathsf{DT}_A = i_A \mathbin{\square} f$. Passing to the complement we get $\mathsf{DT}_A = -\mathsf{DT}_{\complement A} = -(i_{\complement A} \mathbin{\square} f)$ in A. From this the result follows.

Proposition 5.1.6. *Let G be an abelian group with a translation-invariant metric d. Then for any subsets A, B of G we have*

$$(\mathsf{DT}_{A+B})^+ = (\mathsf{DT}_A)^+ \mathbin{\square} i_B = i_A \mathbin{\square} (\mathsf{DT}_B)^+ = (\mathsf{DT}_A)^+ \mathbin{\square} (\mathsf{DT}_B)^+.$$

Proof. We know from the preceding proposition that $(\mathsf{DT}_A)^+ = i_A \mathbin{\square} f$, where $f(x) = d(x,0)$ is the distance from x to the origin. Hence, using freely the associativity and commutativity of infimal convolution as well as the functional equation $f \mathbin{\square} f = f$ (cf. Lemma 5.3.3 below),

$$(\mathsf{DT}_A)^+ \mathbin{\square} (\mathsf{DT}_B)^+ = (i_A \mathbin{\square} f) \mathbin{\square} (i_B \mathbin{\square} f) = (i_A \mathbin{\square} i_B) \mathbin{\square} f = i_{A+B} \mathbin{\square} f = (\mathsf{DT}_{A+B})^+.$$

Also

$$(\mathsf{DT}_A)^+ \mathbin{\square} i_B = (i_A \mathbin{\square} f) \mathbin{\square} i_B = (i_A \mathbin{\square} i_B) \mathbin{\square} f = i_{A+B} \mathbin{\square} f = (\mathsf{DT}_{A+B})^+.$$

5.2. Distance transforms and sublevel sets

The *sublevel sets* of a function $f\colon X \to [-\infty, +\infty]$ are the sets of the form

$$\{x \in X; f(x) < s\} \text{ or } \{x \in X; f(x) \leqslant s\}$$

for some element s of $[-\infty, +\infty]$. For brevity we shall denote them by $\{f < s\}$ rather than $\{x \in X; f(x) < s\}$ when no misunderstanding seems possible.

Lemma 5.2.1. *If X is a metric space with metric d, and DT_A is the distance transform of a subset A of X calculated with the use of d, then the closure, interior and boundary of A can all be recovered from knowledge of the sublevel sets of DT_A:*

$$\overline{A} = \{\mathsf{DT}_A \leqslant 0\}, \qquad A^\circ = \{\mathsf{DT}_A < 0\}, \qquad \partial A = \{\mathsf{DT}_A = 0\}.$$

Moreover $\mathsf{DT}_{\overline{A}} = \mathsf{DT}_A$ in $\complement A$ and $\mathsf{DT}_{A^\circ} = \mathsf{DT}_A$ in A.

Proof. To prove the equality $\overline{A} = \{x; \mathsf{DT}_A(x) \leqslant 0\}$, first note that if $x \in A$, then $\mathsf{DT}_A(x) \leqslant 0$ by definition. If on the other hand $x \notin A$, then $x \in \overline{A}$ if and only if there are points $y \in A$ such that $d(x,y)$ is arbitrarily small, which happens if and only if $\mathsf{DT}_A(x) = 0$. This proves the first equality.

The second equality in the statement follows by passing to the complement, and the third by taking the set-theoretical difference $\overline{A} \smallsetminus A^\circ = \partial A$.

If A is any subset of $\mathbf{R}^n$ satisfying $B_<(c,r) \subset A \subset B_\leqslant(c,r)$, where $r > 0$ and we use the distance $d(x,y) = \|x-y\|$ defined by some norm $\|\cdot\|$ on $\mathbf{R}^n$, then $\mathsf{DT}_A(x) = \|x-c\|-r$ provided $r > 0$. This simple example shows that we cannot expect to recover A exactly from DT_A; we have to be content with its interior and closure. However, if $X = \mathbf{Z}^n$, then the topology induced by a norm in $\mathbf{R}^n$ is the discrete topology, so that, for any set A,

$$\overline{A} = A^\circ = A = \{\mathsf{DT}_A < 0\} = \{\mathsf{DT}_A \leqslant 0\}.$$

The boundary is empty and DT_A never takes the value zero.

Proposition 5.2.2. *Let G be an abelian group with a translation-invariant metric d, and let A be an arbitrary subset of G. Then for all positive numbers r and ε we have*

$$\{\mathsf{DT}_A < r\} = A + B_<(0,r) \subset A + B_\leqslant(0,r) \subset \{\mathsf{DT}_A \leqslant r\} \subset \{\mathsf{DT}_A < r+\varepsilon\};$$

and

$$\{\mathsf{DT}_A \leqslant -r\} = A \ominus B_<(0,r) \supset A \ominus B_\leqslant(0,r) \supset \{\mathsf{DT}_A < -r\} \supset \{\mathsf{DT}_A \leqslant -r-\varepsilon\}.$$

This is easy; we omit the proof. It is enough to prove the first chain of inclusions; the second follows by duality.

Give examples to show that where an inclusion sign is written, the inclusion may be strict.

The dilations by the balls $B = B_<(0,r)$, $A + B_<(0,r) = \delta_B(A)$, thus determine the strict sublevel sets of DT_A for positive values; similarly for the erosions $A \ominus B_<(0,r) = \varepsilon_B(A)$ and the nonstrict sublevel sets of DT_A for negative values.

Proposition 5.2.3. *Let G be an abelian group and $f_j \colon G \mapsto [-\infty, +\infty]$, $j = 1,2$, two arbitrary functions defined on G. Define $f_3 = f_1 \mathbin{\square} f_2$. Then for all real numbers r_1, r_2 and $r_3 = r_1 + r_2$ we have*

$$\{f_1 < r_1\} + \{f_2 < r_2\} \subset \{f_1 < r_1\} + \{f_2 \leqslant r_2\} \subset \{f_3 < r_3\}; \tag{5.2.1}$$

$$\{f_1 \leqslant r_1\} + \{f_2 \leqslant r_2\} \subset \{f_3 \leqslant r_3\}. \tag{5.2.2}$$

Moreover, for any real number r_3 we have

$$\bigcup_{r_1 \in \mathbf{R}} \big(\{f_1 < r_1\} + \{f_2 < r_3 - r_1\}\big) = \{f_3 < r_3\}. \tag{5.2.3}$$

Proof. We always have $(f_1 \mathbin{\square} f_2)(x_1 + x_2) \leqslant f_1(x_1) \dotplus f_2(x_2)$, so if $f_1(x_1) < r_1$ and $f_2(x_2) \leqslant r_2$, then $f_3(x_1 + x_2) < r_1 + r_2 = r_3$. This proves (5.2.1). Similarly if $f_1(x_1) \leqslant r_1$.

Formula (5.2.3) follows from (3.2.10). We can also prove it directly as follows. Assume that $f_3(x_3) < r_3$. Then there exist two points x_1 and x_2 with $x_1 + x_2 = x_3$ such that $f_1(x_1) + f_2(x_2) < r_3$. It is now possible to find a number r_1 such that $f_1(x_1) < r_1$ and $f_2(x_2) < r_3 - r_1$. This proves one inclusion in (5.2.3); the other follows from (5.2.1).

When the functions f_j are distance transforms and r_3 is positive we can improve (5.2.3):

Proposition 5.2.4. *Let G be an abelian group equipped with a translation-invariant metric, and let A_j, $j = 1, 2$, be two subsets. Then their distance transforms $f_j = \mathsf{DT}_{A_j}$ satisfy*

$$\{f_1 \mathbin{\square} f_2 < r\} = \big(\{f_1 \leqslant 0\} + \{f_2 < r\}\big) \cup \big(\{f_1 < r\} + \{f_2 \leqslant 0\}\big) = A_1 + A_2 + B_<(0, r)$$

for all positive r.

Proof. That $\{f_1 \mathbin{\square} f_2 < r\}$ contains $\{f_1 \leqslant 0\} \cup \{f_2 < r\}$ follows from Proposition 5.2.3.

To prove the inclusion in the other direction, let us assume that $(f_1 \mathbin{\square} f_2)(x_3) < r$. Then there exists x_1, x_2 such that $x_1 + x_2 = x_3$ and $f_1(x_1) + f_2(x_2) < r$. If both $f_1(x_1)$ and $f_2(x_2)$ happen to be nonpositive, we are done. The case when $f_1(x_1)$ or $f_2(x_2)$ is positive remains to be considered. Assume first that $f_1(x_1)$ is positive. Then $x_1 \notin A_1$ and we know that, to any given positive ε, there exists a point $y_1 \in A_1$ such that $d(x_1, y_1) < f_1(x_1) + \varepsilon$. We may choose $\varepsilon = r - f_1(x_1) - f_2(x_2)$. Define $y_2 = x_1 + x_2 - y_1$. Then $f_1(y_1) \leqslant 0$ and the Lipschitz continuity of f_2 implies that

$$f_2(y_2) \leqslant f_2(x_2) + d(y_2, x_2) < f_1(x_1) + f_2(x_2) + \varepsilon = r.$$

(Note that $y_2 - x_2 = x_1 - y_1$, so that $d(y_2, x_2) = d(y_1, x_1) < f_1(x_1) + \varepsilon$.) Hence $x_3 = x_1 + x_2 = y_1 + y_2 \in \{f_1 \leqslant 0\} + \{f_2 < r\}$. The case $f_2(x_2) > 0$ is similar.

To prove the last equality we note that $\{f_1 \leqslant 0\}$ is equal to the closure $\overline{A_1}$ of A_1 and that $\{f_2 < r\} = A_2 + B_<(0, r)$. However $\overline{A} + B = A + B$ if B is open, so both terms in the union simplify to $A_1 + A_2 + B_<(0, r)$.

5.3. Chamfer distances

While the Euclidean distance is easy to visualize geometrically, it has certain drawbacks when it comes to calculations: we need to keep in memory a vector rather than a scalar at each pixel; we need more operations per pixel; and, perhaps most importantly, the Euclidean distance is more difficult to use for various morphological operations, such as skeletonizing, than for instance the city-block distance; see Borgefors (1994). For a study of the computation of the Euclidean distance transform in any dimension, see Ragnemalm (1993).

In the case of the city-block (l^1) and chessboard (l^∞) distances, one first defines the distances between neighboring pixels; we shall call them, following Starovoitov (1995:501), *prime distances.* Then the distance between any two pixels is defined by following a path and taking as the distance the minimum over all admissible paths of the sum of the prime distances. As an example, for the city-block distance the admissible paths consists of horizontal and vertical moves only, and the prime distance between two pixels which share a side is declared to be one. Thus the distance is calculated successively from neighboring pixels, which is convenient both for sequential and parallel computation. This is impossible for the Euclidean distance in spaces of dimension two or more.

It turns out that many metrics used in image analysis are conveniently defined from the prime distances by infimal convolution over all grid points. We shall now explain this, following Kiselman (1996).

The following result is well known and easy to prove.

Lemma 5.3.1. *Any translation-invariant distance d on an abelian group G defines a function $f(x) = d(x, 0)$ on X which is positive definite:*

$$f(x) \geqslant 0 \text{ with equality precisely when } x = 0; \tag{5.3.1}$$

and symmetric:

$$f(-x) = f(x) \text{ for all } x \in X. \tag{5.3.2}$$

Conversely, a function f which satisfies (5.3.1) and (5.3.2) defines a distance $d(x, y) = f(x - y)$.

Lemma 5.3.2. *Let d be a translation-invariant distance on an abelian group G and f a function on G related to d as in Lemma 5.3.1. Then d is a metric if and only if f is subadditive:*

$$f(x + y) \leqslant f(x) + f(y) \text{ for all } x, y \in X. \tag{5.3.3}$$

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,

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

proving the triangle inequality.

In the definition of an infimal convolution the infimum operator acts over an infinite set of points, and therefore sometimes cannot be computed in finitely many steps. However, there are many situations where the infimum is in fact a minimum over a finite set. One such case is when f is bounded from below and g is coercive in the strong sense that all sublevel sets $\{y;\ g(y) \leqslant a\}$, $a \in \mathbf{R}$, are finite. Then in particular the sublevel set $\{y;\ g(y) \leqslant (f \,\square\, g)(x) + 1 - \inf f\}$ is finite for every x, and it is enough to search for a minimizing y in that set. Even simpler is the case when g is less than

$+\infty$ in a finite set P only. Then the infimal convolution with any function f is equal to the minimum

$$(f \mathbin{\Box} g)(x) = \min_{y \in P} \bigl(f(x-y) \dotplus g(y)\bigr), \qquad x \in G.$$

This is indeed the case for the distances we shall consider: here P is a small set around the origin where the prime distances are defined.

We have seen that subadditive functions are important when it comes to defining metrics (Lemma 5.3.2). Therefore it is of interest to know that subadditivity can be characterized using infimal convolution:

Lemma 5.3.3. *A function f on an abelian group is subadditive in the sense of (5.3.3) if and only if it satisfies the inequality $f \mathbin{\Box} f \geqslant f$. If $f(0) = 0$, this is equivalent to the equation $f \mathbin{\Box} f = f$.*

Proof. If f is subadditive we have $f(x-y) \dotplus f(y) \geqslant f(x)$, so taking the infimum over y gives $(f \mathbin{\Box} f)(x) \geqslant f(x)$. Conversely, $f(x) \dotplus f(y) \geqslant (f \mathbin{\Box} f)(x+y)$ for all x, y, so $f \mathbin{\Box} f \geqslant f$ implies subadditivity. Finally, we always have $(f \mathbin{\Box} f)(x) \leqslant f(x) \dotplus f(0)$, so if $f(0) = 0$ it follows that $f \mathbin{\Box} f \leqslant f$.

Infimal convolution is a commutative and associative operation on functions (see Proposition 3.2.1), so we can write iterated convolutions as $f \mathbin{\Box} g \mathbin{\Box} h$ without using parentheses. A k-fold convolution can be defined by

$$(f_1 \mathbin{\Box} \cdots \mathbin{\Box} f_k)(x) = \inf \sum_{j=1}^{k} f_j(x^j), \qquad x \in G, \tag{5.3.4}$$

where the infimum is over all choices of elements $x^j \in G$ such that $x^1 + \cdots + x^k = x$, and with the understanding that the sum receives the value $+\infty$ as soon as one of the terms has that value, even in the presence of a value $-\infty$. In (5.3.4) it is natural to think of a path leading from 0 to x consisting of segments $[0, x^1]$, $[x^1, x^1 + x^2]$, $\ldots$, $[x^1 + \cdots + x^{k-1}, x]$; if $G = \mathbf{Z}^2$ this path can be realized in $\mathbf{R}^2$.

If A is a subset of an abelian group G, we shall write $\mathbf{N} \cdot A$ for the semigroup generated by A:

$$\mathbf{N} \cdot A = \left\{\textstyle\sum m_i a_i;\ m_i \in \mathbf{N}, a_i \in A\right\},$$

where all but finitely many of the m_i are zero. Similarly, we shall write $\mathbf{Z} \cdot A$ for the group generated by A:

$$\mathbf{Z} \cdot A = \left\{\textstyle\sum m_i a_i;\ m_i \in \mathbf{Z}, a_i \in A\right\}.$$

If A is symmetric, $A = -A$, then of course $\mathbf{Z} \cdot A = \mathbf{N} \cdot A$.

It seems plausible that if a repeated convolution $F \mathbin{\Box} F \mathbin{\Box} \cdots \mathbin{\Box} F$ has a limit f as the number of terms tends to infinity, then this limit will satisfy the equation $f \mathbin{\Box} f = f$. This is actually so under very general hypotheses:

Theorem 5.3.4. *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 $F_1 = F$, $F_j = F_{j-1} \mathbin{\Box} F$, $j = 2, 3, \ldots$, in other words, F_j is the infimal convolution of j terms all equal to F. Then the sequence $(F_j)_j$ is decreasing and its limit* $\lim F_j = f \geqslant 0$ *is subadditive.*

Moreover $\operatorname{dom} f = \mathbf{N} \cdot \operatorname{dom} F$*, i.e.,* f *is finite precisely in the semigroup generated by* $\operatorname{dom} F$*.*

Remark. It is easy to prove that f is the largest subadditive minorant of F.

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

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

Next we shall prove that $f(x+y) \leqslant f(x) + f(y)$. If one of $f(x), f(y)$ is equal to $+\infty$ there is nothing to prove, so let x, y be given with $f(x), f(y) < +\infty$ and fix a positive number ε. Then there exist numbers j, k such that $F_j(x) \leqslant f(x) + \varepsilon$ and $F_k(y) \leqslant f(y) + \varepsilon$. By associativity $F_{j+k} = F_j \,\Box\, F_k$, 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. Finally, the statement about $\operatorname{dom} g$ is an easy consequence of (3.2.6).

Theorem 5.3.5. *With* F *as in Theorem 5.3.4, assume in addition that there is a translation-invariant metric* d_1 *on* G *such that* $F(x) \geqslant d_1(x, 0)$ *for all* $x \in G$*. Then the limit* f *of the sequence* F_j *also satisfies this inequality,* $f(x) \geqslant d_1(x, 0)$*, so that it is positive definite. If* F *is symmetric,* f *is also symmetric and defines a metric* $d(x, y) = f(x-y) \geqslant d_1(x, y)$ *on the subgroup* $\mathbf{Z} \cdot P = \mathbf{N} \cdot P$ *of* G *generated by* $P = \operatorname{dom} F$*.*

Proof. Define $H(x) = d_1(x, 0)$ and let H_j be the infimal convolution of j terms equal to H. From Lemmas 5.3.2 and 5.3.3 it follows that $H \,\Box\, H = H$ and so all H_j are equal to H. Therefore $F \geqslant H$ implies $F_j \geqslant H$ and also the limit f must satisfy $f \geqslant H$. This proves the theorem.

When applying this theorem we could for instance let d_1 be εd_0, where ε is a small positive number and d_0 is the discrete metric defined by (5.1.4). In $\mathbf{Z}^n$ we can also use $d_1(x, y) = \varepsilon \|x - y\|$ for any norm on $\mathbf{R}^n$.

Corollary 5.3.6. *Let* P *be a finite set in an abelian group* G *containing the origin, and let* F *be a function on* G *with* $F(0) = 0$*, taking the value* $+\infty$ *outside* P *and finite positive values at all points in* $P \smallsetminus \{0\}$*. Then* $f = \lim F_j$ *is a positive definite subadditive function. If* P *is symmetric and* $F(-x) = F(x)$*, then* f *defines a metric on the subgroup* $\mathbf{Z} \cdot P = \mathbf{N} \cdot P$ *of* G *generated by* P*.*

Proof. Since P is finite, there is a positive number ε such that $F(x) \geqslant \varepsilon$ for all $x \in P$ except $x = 0$. Thus $F(x) \geqslant \varepsilon d_0(x, 0)$, where d_0 is the discrete metric defined by (5.1.4). We can now apply the theorem.

Definition 5.3.7. *Let us say that a metric* $d(x, y) = f(x - y)$ *is a* chamfer distance, *or* finitely generated *if it is constructed as in Corollary 5.3.6.*

It is easy to prove that the Euclidean metric $d(x,y)=\sqrt{\sum(x_j-y_j)^2}$ on $\mathbf{Z}^n$ is a chamfer distance if and only if $n \leqslant 1$.

Borgefors (1984:324, 1986:345) calls $F(x-y)$ the *local distances*; Verwer (1991:672) adopted this term. Starovoitov (1995:501) calls F the *prime distance function.* The resulting metric, $(x,y)\mapsto f(x-y)$, was called a *quasi-Euclidean* distance by Montanari (1968) and a *chamfer distance* by Borgefors (1984:326). The term has then been used, e.g., by Verwer (1991:672) and Marchand-Maillet & Sharaiha (2000:21). However, Borgefors herself now prefers the term *weighted distances*, and refers to them as *constructed by chamfering* (Gunilla Borgefors, personal communication 2002-03-22); the latter term is derived from the method of calculating distance transforms by passing a mask twice over the image. Nevertheless, the term *chamfer distance* for the metrics constructed in Corollary 5.3.6 has won acceptance.

It is by no means necessary that f is positively homogeneous in Corollary 5.3.6. In fact, we can let $P=\{0,\pm1,\pm2\}\subset\mathbf{Z}$ and define $F(\pm1)=a$, $F(\pm2)=b$, where a and b are arbitrary positive numbers. If $b\geqslant 2a$, then $f(x)=a|x|$ for all $x\in\mathbf{Z}$, but if $b<2a$, then $f(x)=\frac{1}{2}b|x|$ when x is even, $x=2k$, $k\in\mathbf{Z}$, whereas $f(x)=bk+a>\frac{1}{2}b|x|$ when x is odd, $x=\pm(2k+1)$, $k\in\mathbf{N}$. Nevertheless f is subadditive.

A more interesting example is perhaps this in two dimensions. Let $P=\{x\in\mathbf{Z}^2;\ |x_j|\leqslant 1\}$, and define the prime distances as $F(\pm1,0)=F(0,\pm1)=a>0$, $F(\pm1,\pm1)=b>0$. Then if $b\geqslant a$ we get $f(x_1,0)=a|x_1|$. But if $b<a$, then $f(2,0)=2b<2a$, so that $f(2,0)<2f(1,0)=2a$. In fact, by the definition of infimal convolution, $f(2,0)\leqslant F_2(2,0)\leqslant F(1,1)+F(1,-1)=b+b$. On the other hand, it is not difficult to see that for any k, $F_k(2,0)\geqslant 2b$, so that actually $f(2,0)=2b$. This is because if we take $k\geqslant 2$ nonzero steps to go from the origin to $(2,0)$, the distance assigned to the path is at least $F(x^1)+\cdots+F(x^k)\geqslant kb$.

Several metrics on $\mathbf{Z}^2$ have been studied. When presenting the generating function F defining the prime distances it shall be understood in the sequel that F is invariant under permutation and reflection of the coordinates. Therefore it is enough to define $F(x)$ for $0\leqslant x_2\leqslant x_1$. Also it is understood that $F(0)=0$ in all cases, and that $F(x)=+\infty$ when not mentioned.

Consider first $P=\{x\in\mathbf{Z}^2;\ \sum|x_j|\leqslant 1\}$ and $F(1,0)=1$. Then the corresponding metric is the city-block (l^1) metric, introduced and studied by Rosenfeld & Pfaltz (1966). If instead we let $P=\{x\in\mathbf{Z}^2;\ |x_j|\leqslant 1\}$ and $F(1,0)=F(1,1)=1$, then the metric is the chessboard (l^∞) metric, introduced by Rosenfeld & Pfaltz (1968). Some other metrics that have been studied are modifications of this; to define them, put $F(1,0)=a$ and $F(1,1)=b$. Then the choices $(a,b)=(1,\sqrt{2})$ (Montanari 1968); $(a,b)=(2,3)$ (Hilditch & Rutovitz 1969); and $(a,b)=(3,4)$ (Borgefors 1984) have all been studied. Next we can increase the size of the neighborhood where prime distances are defined to include the knight's move $(2,1)$ as an element of P. The distance defined by this move only has been studied by Das & Chatterji (1988). It seems more natural, however, to allow also $(1,0)$ and $(1,1)$ in P. Then a very good choice under certain criteria is $F(1,0)=5$, $F(1,1)=7$, and $F(2,1)=11$ (the 5-7-11 weighted distance). This distance was proposed and studied by Borgefors (1986).

We always have $f\leqslant F$, and it may happen that $f(x)<F(x)$ for some pixel $x\in P$. Let for instance $F(1,0)=a$, $F(2,1)=c$, and extend F by reflection and permutation

of the coordinates. Then

$$f(1,0) \leqslant F_3(1,0) \leqslant F(2,1) + F(1,-2) + F(-2,1) = 3c,$$

so if $3c < a$ we get $f(1,0) \leqslant 3c < a = F(1,0)$. This is undesirable, because we expect the prime distance originally defined between the origin and $(1,0) \in P$ to survive and to be equal to the distance defined by the minimum over all paths. It is therefore natural to require that $f = F$ everywhere in P.

Since $f(x)$ is the limit of an infinite sequence $F_j(x)$, it is reassuring to know that this sequence is in fact stationary in the cases of interest here. It is easy to explicitly give an index j such that $F_j(x)$ is equal to the limit $f(x)$:

Proposition 5.3.8. *Let F be as in Corollary 5.3.6. Then the sequence (F_j) is pointwise stationary, i.e., for every $x \in G$ there is an index $s(x)$ such that $F_j(x) = f(x)$ for all $j \geqslant s(x)$.*

Proof. We first note that by hypothesis there are two constants $c > 0$ and C such that

$$c \leqslant F(p) \leqslant C, \qquad p \in P \smallsetminus \{0\}.$$

If $x \notin \mathbf{N} \cdot P$, then $f(x) = F_j(x) = +\infty$ for all j. If $x \in \mathbf{N} \cdot P$, then there is an index $m(x)$ such that $x \in P + \cdots + P$ with $m(x)$ terms, and $F_{m(x)}(x) \leqslant m(x)C$. For every $j \geqslant m(x)$ there are elements $y^1, y^2, \ldots, y^j$ in P such that $x = y^1 + \cdots + y^j$ and $F_j(x) = F(y^1) + \cdots + F(y^j)$ (cf. (5.3.4)). Since $F \geqslant c$ in $P \smallsetminus \{0\}$ we get

$$Cm(x) \geqslant F_{m(x)}(x) \geqslant F_j(x) \geqslant qc,$$

where q is the number of indices i such that $y^i \neq 0$. Hence the number q of nonzero terms in any representation of $F_j(x)$ with $j \geqslant m(x)$ can never be larger than $Cm(x)/c$. Now define $s(x) = \lfloor Cm(x)/c \rfloor$. If $j \geqslant s(x)$, then $j \geqslant q$ and we have $F_j(x) = F_q(x)$, since, in the formation of $F_j(x)$, at most q of the terms in a sum $x = y^1 + \cdots + y^j$ can be nonzero if the value of the sum $F(y^1) + \cdots + F(y^j)$ shall come down to $F_q(x)$. This means that the sequence $F_j(x)$ is stationary starting with $s(x)$.

Remark. It is not true that $F_j(x) = F_{j-1}(x)$ implies that $F_k(x) = F_j(x)$ for all $k \geqslant j$, so $F_j(x) = F_{j-1}(x)$ at a particular point x is not a sufficient criterion. For example, we may define $F(\pm 1) = 1$ and $F(\pm 100) = 101$. Then $F_j(100) = 101$ for $j = 1, \ldots, 99$ but $F(100) = 100$.

Remark. The numbers C and c are structural constants of the prime distances $F(p)$, $p \in P$, and can be taken as

$$C = \max_{p \in P} F(p) \text{ and } c = \min_{p \in P \smallsetminus \{0\}} F(p).$$

The number $m(x)$ can be easily estimated in the most common applications. For instance, for the l^∞ metric we can take $C = c = 1$ and $m(x) = \|x\|_\infty$. Similarly for the

l^1 metric we can take $C = c = 1$ and $m(x) = \|x\|_1$. It is therefore easy to estimate the index $s(x)$.

Corollary 5.3.9. *The positive part of a distance transform is a limit* $(\mathsf{DT}_A)^+ = \lim(i_A \,\Box\, F \,\Box\, F \,\Box\, \cdots \,\Box\, F)$*, where the number of terms tends to infinity. This formula can be used in actual calculations: starting from* $g_0 = i_A$ *one calculates* $g_j(x) = (g_{j-1} \,\Box\, F)(x)$ *and stops when the criterion of Proposition 5.3.8 is satisfied.*

5.4. Comparing distances

The l^1 (city-block)and l^∞ (chessboard) metrics in $\mathbf{R}^2$ are translation invariant but not rotation invariant. (In the plane a rotation can distort distances by a factor of up to $\sqrt{2}$; in higher dimensions more.) The Euclidean metric is rotation invariant, and it is desirable to construct a chamfer distance in $\mathbf{Z}^n$ which is reasonably close to being rotation invariant. There are many studies on the problem of defining an optimal distance in a given family of finitely generated distances. Of course the property of being optimal depends on the criteria employed; beauty is in the eye of the beholder. A basic problem is how to measure deviation: we may ask how far the quotient of two quantities is from 1, alternatively how far their difference is from 0. In this section we shall look briefly into this problem and describe four methods of comparing two nonnegative functions; two of these have been studied earlier.

It is natural to measure the deviation of a function $f\colon X \to [0, +\infty[$ from a given nonnegative function g defined on the same set by the smallest constant $C \in [0, +\infty]$ such the inequalities $f(x) \leqslant Cg(x)$ and $g(x) \leqslant Cf(x)$ hold for all $x \in X$. We introduce a notation for this constant,

$$\Lambda(f, g) = \max\left(\sup_{x\in X} \frac{f(x)}{g(x)}, \sup_{x\in X} \frac{g(x)}{f(x)}\right), \tag{5.4.1}$$

where the supremum is taken over all points in the common domain of definition, and where we count $0/0$ as 0 and $t/0$ as $+\infty$ if $t > 0$ (this is to allow for zeros; $\Lambda(f, g)$ is finite only if the two functions have the same zero set). It is noteworthy that $\log \Lambda(f, g) = \|\log f - \log g\|_\infty$ is a distance on a suitable space of functions; in particular it is symmetric.

If f satisfies an inequality $C_1 \leqslant f(x)/g(x) \leqslant C_2$, then a slightly modified function, viz. $f_1 = f/\sqrt{C_1C_2}$, satisfies $\Lambda(f_1, g) \leqslant \sqrt{C_2/C_1}$.

Another, closely related measure of the deviation was studied by Verwer (1991). He used the functional

$$\Lambda'(f, g) = \sup_{x\in X} \left|\frac{f(x)}{g(x)} - 1\right|. \tag{5.4.2}$$

However, one might just as well consider $\Lambda'(g, f)$. It is readily seen that $\Lambda'(f, g) = \Lambda(f, g) - 1$ when $f \geqslant g$, and that $\Lambda'(f, g) = 1 - 1/\Lambda(f, g)$ when $f \leqslant g$. In general, $\Lambda'(f, g)$, $\Lambda'(g, f)$ as well as $\log \Lambda(f, g)$ lie between two limits,

$$1 - \frac{1}{\Lambda(f, g)} \leqslant \Lambda'(f, g), \Lambda'(g, f), \log \Lambda(f, g) \leqslant \Lambda(f, g) - 1,$$

where we have put in also the well-known inequality $1 - 1/t \leqslant \log t \leqslant t - 1$. In particular, $|\Lambda'(f,g) - \log\Lambda(f,g)| \leqslant (\Lambda(f,g) - 1)^2/\Lambda(f,g)$; the same estimate holds of course for $\Lambda'(g,f)$. We may also note that, although Λ' is not symmetric, it is approximately symmetric when f and g are close, and there is an estimate

$$\frac{\Lambda'(f,g)}{\Lambda(f,g)} \leqslant \Lambda'(g,f) \leqslant \Lambda(f,g)\Lambda'(f,g),$$

which, by the way, may be written as

$$\Lambda(\Lambda'(f,g), \Lambda'(g,f)) \leqslant \Lambda(f,g).$$

When f and g are reasonably close, $\Lambda'(f,g) \approx \Lambda'(g,f) \approx \log\Lambda(f,g)$. For many purposes either one may be used. Note, however, that $\Lambda(f,g)$ has better functional properties than $\Lambda'(f,g)$. In particular, as already noted, $\log\Lambda(f,g)$ is a metric, whereas $\Lambda'(f,g)$ does not satisfy the triangle inequality and is not even symmetric.

We note that for every pair (f,g) of functions there are constants c_0, c_1, and c_2 such that, respectively, $\Lambda(c_0 f, g)$, $\Lambda'(c_1 f, g)$ and $\Lambda'(g, c_2 f)$ are minimal. It is easy to see that c_0 is the geometric mean of c_1 and c_2.

In particular we shall compare the chamfer distances with the Euclidean norm $g = \|\cdot\|_2$. In the exercises in this chapter we let the reader determine or estimate the deviation from the Euclidean norm of some well-known finitely generated metrics. If the prime vectors are $(\pm1, 0)$, $(0, \pm1)$ and $(\pm1, \pm1)$ with prime distances a and b respectively, we note that the optimal prime distances for both Λ and Λ' are related by $b = a\sqrt{2}$, but that the actual values are slightly different. For $\Lambda(f, \|\cdot\|_2)$, the optimal choice is

(5.4.3)

$$a = a_0 = \sqrt[4]{\frac{2+\sqrt{2}}{4}} \approx 0.961186523, \qquad b = b_0 = a_0\sqrt{2} = \sqrt[4]{2+\sqrt{2}} \approx 1.359323017,$$

whereas Verwer (1991:676) found the optimal choice for $\Lambda'(f, \|\cdot\|_2)$ to be

$$a = a_1 \approx 0.9604 \text{ and } b = b_1 = a_1\sqrt{2} \approx 1.3583.$$

The exact values are

$$a_1 = \frac{1}{\frac{1}{2} + \sqrt{1 - 1/\sqrt{2}}} \text{ and } b_1 = a_1\sqrt{2}. \tag{5.4.4}$$

One can calculate also the optimal choice for $\Lambda'(\|\cdot\|_2, f)$, which is

$$a = a_2 = \tfrac{1}{2} + \frac{1}{4}\sqrt{2+\sqrt{2}} \approx 0.961939766 \text{ and } b = b_2 = a_2\sqrt{2} \approx 1.3603882. \tag{5.4.5}$$

In this case the optimality has a clear geometrical meaning: the vertices of the octagon protrude as much outside the disk as the midpoints of the edges go into the disk. As we can expect from an earlier remark, $a_0 = \sqrt{a_1 a_2}$.

In her pioneering work on chamfer distances (1984, 1986), Gunilla Borgefors used instead the functional

$$\Lambda'_{2,\infty}(f) = \sup_{x\in\mathbf{Z}^2} \left| \frac{f(x)}{\|x\|_2} - 1 \right| \frac{\|x\|_2}{\|x\|_\infty} = \sup_{x\in\mathbf{Z}^2} \frac{\bigl|f(x) - \|x\|_2\bigr|}{\|x\|_\infty}, \tag{5.4.6}$$

and she determined the optimal distances for this measure as follows:

$$\Lambda'_{2,\infty}(f) = \tfrac{1}{2}\sqrt{2\sqrt{2}-2} - \tfrac{1}{2} \approx 0.04491,$$

attained for
(5.4.7)

$$a = a_3 = \tfrac{1}{2}\sqrt{2\sqrt{2}-2} + \tfrac{1}{2} \approx 0.95509 \text{ and } b = b_3 = \sqrt{2} - \tfrac{1}{2} + \tfrac{1}{2}\sqrt{2\sqrt{2}-2} \approx 1.36930;$$

note that $b_3 > a_3\sqrt{2}$ here (1986:351). She also determined the optimal values for the functional $\Lambda'_{2,\infty}$ when a is restricted to be 1 and b is free to vary; in this case $\Lambda'_{2,\infty}(f) = 1/\sqrt{2} - \sqrt{\sqrt{2}-1} \approx 0.06$ and is attained for

$$a = a_4 = 1 \text{ and } b = b_4 = 1/\sqrt{2} - \sqrt{\sqrt{2}-1} \approx 1.351; \tag{5.4.8}$$

here $b_4 < a_4\sqrt{2}$ (1984:327).

5.5. The calculus of balls

In any metric space the inequality $d(a,b) + r \leqslant s$ implies that the open ball $B_<(a,r)$ is contained in $B_<(b,s)$ and also that $B_\leqslant(a,r) \subset B_\leqslant(b,s)$. In a normed vector space of dimension at least one and equipped with the distance $d(x,y) = \|x-y\|$ defined by the norm, the converse implications hold, provided $r > 0$ ($r \geqslant 0$ for closed balls). In particular, two open balls $B_<(a,r)$ and $B_<(b,s)$ with $r > 0$ are equal if and only if $a = b$ and $r = s$ ($r \geqslant 0$ suffices for closed balls). But in a general metric space very little can be said: from $B_\leqslant(a,r) \subset B_\leqslant(b,s)$ we can only deduce that $d(a,b) \leqslant s$, assuming that $r \geqslant 0$.

We note, however, that if, in an abelian group with a translation-invariant metric, $B_\leqslant(a,r)$ is contained in $B_\leqslant(b,s)$, $s \leqslant r$, and $r \geqslant 0$, then $a - b$ belongs to a bounded subgroup. In the most common applications, like normed spaces and groups with a chamfer distance, the only bounded subgroup is $\{0\}$, thus $a = b$.

The mappings

$$X \times \mathbf{R} \ni (a,r) \mapsto B_\leqslant(a,r), B_<(a,r) \in \mathscr{P}(X)$$

are in general far from injective and induce complicated equivalence relations in $X \times \mathbf{R}$. Similarly, the inclusion relations $B_\leqslant(a,r) \subset B_\leqslant(b,s)$, $B_<(a,r) \subset B_<(b,s)$ induce preorders in $X \times \mathbf{R}$ (see (2.1.1) and (2.1.2)). We shall now study these relations.

In any metric space X with metric d we fix a point c in X and define two functions which are of interest when defining balls:

$$\rho(r) = \sup_{x\in X}\bigl(d(x,c); d(x,c) \leqslant r\bigr) \text{ and } \sigma(r) = \inf_{x\in X}\bigl(d(x,c); d(x,c) \geqslant r\bigr), \quad r \in [-\infty, +\infty].$$

The functions depend of course in general on the choice of c, but in most applications we shall let X be an abelian group and assume that d is translation invariant. Then the choice of c is immaterial.

The functions

$$\rho_-(r) = \sup_{s<r} \rho(s) = \sup_{x\in X} \big(d(x,c); d(x,c) < r\big), \qquad r \in [-\infty, +\infty],$$

and

$$\sigma_+(r) = \inf_{s>r} \sigma(s) = \inf_{x\in X} \big(d(x,c); d(x,c) > r\big), \qquad r \in [-\infty, +\infty],$$

are also of interest. It is clear that

$$\rho_- \leqslant \rho \leqslant \mathsf{Id}_{[-\infty,+\infty]} \leqslant \sigma \leqslant \sigma_+.$$

On the negative half-axis we have $\rho(r) = -\infty$ for $-\infty \leqslant r < 0$ and $\sigma(r) = 0$ for $-\infty \leqslant r \leqslant 0$.

We have $B_\leqslant(c,r) = B_\leqslant(c,\rho(r))$ and $B_<(c,r) = B_<(c,\sigma(r))$. The open interval $]\rho(r),\sigma(r)[$ contains no distances $d(c,x)$—a sphere with center at c and radius in that interval is empty. The functions ρ and σ solve the uniqueness problem for balls with a common center: $B_\leqslant(c,r) = B_\leqslant(c,s)$ if and only if $\rho(r) = \rho(s)$, and $B_<(c,r) = B_<(c,s)$ if and only if $\sigma(r) = \sigma(s)$.

The mappings $\rho,\sigma\colon [-\infty,+\infty] \to [-\infty,+\infty]$ are idempotent; in fact ρ is an opening and σ is a closing (Definitions 2.2.1 and 2.2.3). By way of contrast, ρ_- and σ_+ are in general not idempotent. However, $\rho\circ\rho_- = \rho_-$.

In an abelian group with a translation-invariant distance, the value of the distance transform of a set A at a point $x \in A$, $\mathsf{DT}_A(x) = -r$, satisfies $r = \sigma(r)$. Similarly for $x \notin A$, we have $\mathsf{DT}_A(x) = r$ with $r = \sigma(r)$. Thus the values of $|\mathsf{DT}_A|$ are always contained in the invariant set Inv_σ of σ.

In a normed vector space of dimension at least one, we always have $\rho(r) = \sigma(r) = r$ for positive r, so there is no need to introduce them in the study of such spaces. In $\mathbf{Z}^n$ equipped with the l^∞ or l^1 metrics we have $\rho(r) = \lfloor r\rfloor$ and $\sigma(r) = \lceil r\rceil$ for positive r.

While the functions ρ and σ solve the problem of comparing two balls with a common center, more hypotheses will be needed if we are to compare successfully balls with different centers. This is why we are led to an analysis of the triangle inequality.

We shall say that a translation-invariant metric $d(x,y) = f(x-y)$ defined on an abelian group is *upper regular for the triangle inequality* if, given any x and y, there is a point $\tilde y$ such that

$$f(\tilde y) = f(y) \text{ and } f(x+\tilde y) = f(x) + f(\tilde y). \tag{5.5.1}$$

We shall say that the distance is *lower regular for the triangle inequality* if given any x and y such that $f(x) \geqslant f(y)$, there is a point $\tilde x$ such that

$$f(\tilde x) = f(x) \text{ and } f(\tilde x) = f(\tilde x - y) + f(y). \tag{5.5.2}$$

Upper regularity means that in the triangle inequality $f(x+y) \leqslant f(x) + f(y)$ we can always find $\tilde y$ at the same distance from the origin as y and turning the inequality into an equality (by raising the left-hand side); lower regularity means that we can lower the right-hand side in the triangle inequality $f(x) \leqslant f(x-y)+f(y)$ without changing the left-hand side by replacing x by $\tilde{x}$—but of course only if the left-hand side is at least as large as $f(y)$.

It is clear that normed spaces are both upper and lower regular for the triangle inequality; we may choose $\tilde{y}$ as a suitable multiple of x if $x \neq 0$ and $\tilde{x}$ as a suitable multiple of y if $y \neq 0$. It is also easy to see that in $\mathbf{Z}^n$ the l^∞ and l^1 distances are upper and lower regular for the triangle inequality.

The chamfer distance in $\mathbf{Z}^2$ with $P = \{0, (\pm 1, 0), (0, \pm 1), (\pm 1, \pm 1)\}$ is upper regular provided $a = F(1,0) \leqslant b = F(1,1)$. If $b \geqslant 2a$, then $f(x) = a\|x\|_1$; if $b = a$, then $f(x) = a\|x\|_\infty$. In the remaining cases, when $a < b < 2a$, the balls are octagonal, and we may argue as follows. By symmetry and reflection of the coordinates it is enough to consider a point $x \in \mathbf{Z}^2$ with $0 \leqslant x_2 \leqslant x_1$. Then x can be written as $m(1,0)+n(1,1)$ for some uniquely determined integers $m, n \geqslant 0$, and $f(x) = ma + nb$. If y is in the same sector, thus satisfying $0 \leqslant y_2 \leqslant y_1$, then $f(x+y) = f(x) + f(y)$, so we may take $\tilde{y} = y$. If y is in one of the other seven sectors we take $\tilde{y} = (|y_1|, |y_2|)$ or $(|y_2|, |y_1|)$ so that $\tilde{y}$ is in the same sector as x and we know that we have equality in the triangle inequality (5.5.1). However, this chamfer distances is not lower regular if $a < b < 2a$ as can be shown by simple examples. It seems therefore that lower regularity is too stringent a criterion.

Theorem 5.5.1. *Let G be an abelian group and $d(x,y) = f(x-y)$ a translation-invariant metric in G. Assume that d is upper regular for the triangle inequality. Then for all $r, s \in \mathbf{R}$, $B_{\leqslant}(a,r)$ is contained in $B_{\leqslant}(b,s)$ if and only if $d(a,b)+\rho(r) \leqslant \rho(s)$.*

Proof. By the definition of ρ, $B_{\leqslant}(a,r) = B_{\leqslant}(a,\rho(r))$. The inequality $d(a,b)+\rho(r) \leqslant \rho(s)$ implies that $B_{\leqslant}(a,\rho(r)) \subset B_{\leqslant}(a,\rho(s))$ and hence that $B_{\leqslant}(a,r) \subset B_{\leqslant}(b,s)$.

For the other implication we argue as follows, assuming that $B_{\leqslant}(0,r)$ is contained in $B_{\leqslant}(b,s)$ (we may take $a = 0$ to simplify notation). Take a point x in the first ball. Then $f(x) \leqslant \rho(r)$. It follows that $f(x-b) \leqslant \rho(s)$. In the triangle inequality $f(x-b) \leqslant f(x)+f(b)$ we can find an element $\tilde{x}$ with the same distance to the origin as x and turning the inequality into an equality: $f(\tilde{x}-b) = f(\tilde{x})+f(b)$. Now $f(\tilde{x}) = f(x)$ can take values as close to $\rho(r)$ as we like, say larger than $\rho(r)-\varepsilon$ for a given positive ε. We get $\rho(s) \geqslant f(\tilde{x}-b) = f(\tilde{x})+f(b) \geqslant \rho(r)-\varepsilon+d(a,b)$, proving the inequality. Note that $f(\tilde{x}-b) \leqslant \rho(s)$ follows from the inequality $f(\tilde{x}) = f(x) \leqslant \rho(r)$ and the inclusion $B_{\leqslant}(0,r) \subset B_{\leqslant}(b,s)$.

Theorem 5.5.2. *Let G be an abelian group and $d(x,y) = f(x-y)$ a translation-invariant metric in G. Assume that d is lower regular for the triangle inequality. Then for all $r, s \in \mathbf{R}$ with $r > 0$, $B_{<}(a,r)$ is contained in $B_{<}(b,s)$ if and only if $d(a,b)+\sigma(r) \leqslant \sigma(s)$.*

Proof. By the definition of σ, $B_{<}(a,r) = B_{<}(a,\sigma(r))$. The inequality $d(a,b)+\sigma(r) \leqslant \sigma(s)$ implies that $B_{<}(a,\sigma(r)) \subset B_{<}(b,\sigma(s))$, and hence that $B_{<}(a,r) \subset B_{<}(b,s)$.

To prove the other implication we assume that $B_{<}(a,r) \subset B_{<}(0,s)$ (we may take $b=0$). This means that $f(x) \geqslant \sigma(s)$ implies $f(x-a) \geqslant \sigma(r)$. For any given $\varepsilon > 0$ we can find x such that $\sigma(s) \leqslant f(x) \leqslant \sigma(s)+\varepsilon$. Then $f(x-a) \geqslant \sigma(r)$ and we obtain the rather useless string of inequalities

$$\sigma(r) \leqslant f(x-a) \geqslant f(x) - f(a) \leqslant \sigma(s) + \varepsilon - f(a).$$

However, since $f(a) < s \leqslant \sigma(s) \leqslant f(x)$ (at this point we need to know that $a \in B_<(a, r)$, thus that r is positive), we can by hypothesis find $\tilde{x}$ such that $f(\tilde{x}) = f(x)$ (thus $f(\tilde{x} - a) \geqslant \sigma(r)$) and turning the triangle inequality $\geqslant$ in the last formula into an equality:

$$\sigma(r) \leqslant f(\tilde{x} - a) = f(\tilde{x}) - f(a) \leqslant \sigma(s) + \varepsilon - f(a).$$

If we now let ε tend to zero we obtain $\sigma(r) \leqslant \sigma(s) - f(a)$, which concludes the proof.

Since chamfer distances are in general not lower regular for the triangle inequality it is desirable to impose only upper regularity. Now upper regularity was used for the inclusion of closed balls, lower regularity for the inclusion of open balls. But, luckily, for chamfer distances all open balls are closed! We shall now study such spaces.

We consider metric spaces X such that the set of all distances $\{d(x, c); x \in X\}$ from a fixed point $c \in X$ is a discrete subset of $\mathbf{R}$. This is equivalent to the statement that the set $\{d(x, c); x \in X, d(x, c) \leqslant t\}$ is finite for every real number t. The main examples are of course the chamfer distances in any abelian group.

If the set $\{d(x, c); x \in X\}$ of all distances from a fixed point c is discrete, the set is denumerable and either finite or equal to a set $\{r_j; j \in \mathbf{N}\}$. In the latter case we shall choose the indices so that $r_0 = 0$ and $r_j < r_{j+1}$ for all $j \in \mathbf{N}$. A sphere $S(c, r)$ is nonempty if and only if $r = r_j$ for some j. The functions ρ and σ defined above satisfy

$$\rho(r) = r_j, \text{ when } r \in [r_j, r_{j+1}[\,, \qquad \rho_-(r) = r_j, \text{ when } r \in \,]r_j, r_{j+1}]\,,$$

and

$$\sigma(r) = r_j, \text{ when } r \in \,]r_{j-1}, r_j]\,, \qquad \sigma_+(r) = r_j, \text{ when } r \in [r_{j-1}, r_j[\,.$$

Proposition 5.5.3. *Let X be a metric space where the set of all distances from a fixed point c is discrete. Then all closed balls with center at c are open and all open balls with center at c are closed. More exactly,*

$$B_\leqslant(c, r) = B_<(c, \sigma_+(r)) \text{ and } B_<(c, r) = B_\leqslant(c, \rho_-(r)).$$

Proof. We consider the case when the set of distances is infinite—easy modifications will take care of the case of finitely many distances. It is convenient in the proof to define $r_{-1} = -\infty$. Given any $r \in \mathbf{R}$ we choose $j \in \mathbf{N}$ so that $r \in [r_{j-1}, r_j[$. Then $B_\leqslant(c, r) = B_<(c, s)$ for all $s \in \,]r_{j-1}, r_j]$. In particular $B_\leqslant(c, r) = B_<(c, r_j) = B_<(c, \sigma_+(r))$.

Similarly, given any $r \in \mathbf{R}$ we choose $j \in \mathbf{N}$ such that $r \in \,]r_{j-1}, r_j]$; then $B_<(c, r) = B_\leqslant(c, s)$ for all $s \in [r_{j-1}, r_j[$. In particular, $B_<(c, r) = B_\leqslant(c, r_{j-1}) = B_\leqslant(c, \rho_-(r))$ for all $r \in \mathbf{R}$.

Theorem 5.5.4. *Let G be an abelian group and $d(x, y) = f(x - y)$ a translation-invariant metric in G. Assume that d is upper regular for the triangle inequality and* *that the set of all distances is discrete. Then for all* $r, s \in \mathbf{R}$*,* $B_<(a, r)$ *is contained in* $B_<(b, s)$ *if and only if* $d(a, b) + \rho_-(r) \leqslant \rho_-(s)$*.*

Proof. We know that the open ball $B_<(a, r)$ is equal to the closed ball $B_\leqslant(a, \rho_-(r))$. Theorem 5.5.1 shows that the inclusion $B_\leqslant(a, \rho_-(r)) \subset B_\leqslant(b, \rho_-(s))$ is equivalent to the inequality $d(a, b) + \rho(\rho_-(r)) \leqslant \rho(\rho_-(s))$. However, we always have $\rho(\rho_-(r)) = \rho_-(r)$. This completes the proof.

5.6. Distance transforms in normed vector spaces

In this section we shall calculate the distance transforms of some subsets of normed spaces. In such spaces we shall always use the metric defined by the given norm of the space, $d(x, y) = \|x - y\|$.

The space $\mathbf{R}^n$ of all n-tuples can be normed by the l^p*-norm* $\|\cdot\|_p$, $1 \leqslant p \leqslant +\infty$, which is defined by

$$\|x\|_p = \Big(\sum |x_j|^p\Big)^{1/p}, \qquad x = (x_1, \ \dots, x_n) \in \mathbf{R}^n. \tag{5.6.1}$$

When $p = +\infty$ this has to be interpreted as a limit. More explicitly one defines

$$\|x\|_\infty = \max_j |x_j|, \qquad x \in \mathbf{R}^n. \tag{5.6.2}$$

For any normed vector space E we consider its *dual* E', consisting of all continuous linear forms on E. On the dual we define the norm *dual* to $\|\cdot\|$ by $\|\xi\|' = \sup_{\|x\|\leqslant 1} |\xi(x)|$ for $\xi \in E'$. It follows that $|\xi(x)| \leqslant \|\xi\|'\|x\|$ for all $x \in E$ and all $\xi \in E'$.

When $E = \mathbf{R}^n$, we may identify also E' with $\mathbf{R}^n$, and the evaluation of ξ at the point x, i.e., the number $\xi(x)$, is then the inner product $\xi \cdot x$. The Euclidean norm $\|\cdot\|_2$ is dual to itself:

$$\|\xi\|_2' = \sup_{\|x\|_2\leqslant 1} \xi(x) = \|\xi\|_2 = \sqrt{\sum \xi_j^2}.$$

It is not difficult to prove that the norm dual to $\|\cdot\|_1$ is $\|\cdot\|_\infty$ and vice versa. More generally, one can prove that the norm dual to $\|\cdot\|_p$ is $\|\cdot\|_q$, where $q = p/(p-1)$, $1 < p < +\infty$, with a natural interpretation also when $p = 1, +\infty$. This statement follows from Hölder's inequality and its converse.

In all what follows we may take E as $\mathbf{R}^n$ with one of these norms. However, the more general statements are really not more difficult to prove.

In particular we shall look at the case of a convex subset A of a normed vector space. It is easy to see that the distance transform of a ball $B = B_\leqslant(a, r)$ is $\mathsf{DT}_B(x) = \|x - c\| - r$. Another simple convex set is a half-space, and we shall now determine its distance transform:

Proposition 5.6.1. *Let* Y *be a closed half-space in a normed space* E*, defined by an inequality* $\xi(x - a) \leqslant 0$ *for some continuous linear form* $\xi \in E'$*,* $\xi \neq 0$*. Then its distance transform is* $\mathsf{DT}_Y(x) = \eta(x - a)$*, where* $\eta = \xi/\|\xi\|'$*.*

Proof. We may as well define Y by the inequality $\eta(x - a) \leqslant 0$. We note that $\|\eta\|' = 1$.

By definition $\mathsf{DT}_Y(x) = \inf_{\eta(y)\leqslant\eta(a)} \|x-y\|$ when $x \notin Y$. Without loss of generality we may assume that $x=0$ and $\eta(a)=-1$; this can be achieved by a change of variable. We shall thus have to prove that

$$\inf_{\substack{y\in E\\ \eta(y)\leqslant -1}} \|y\| = 1. \tag{5.6.3}$$

This is a well-known fact. To prove it, we note that $\sup_{\|z\|\leqslant 1}\eta(z)=1$ by the definition of the dual norm, and that, given any ε with $0<\varepsilon<1$, we can take z_ε of norm at most 1 so that $\eta(z_\varepsilon)\geqslant 1-\varepsilon$. Then $y_\varepsilon = -z_\varepsilon/\eta(z_\varepsilon)$ satisfies $\eta(y_\varepsilon)=-1$ and $\|y_\varepsilon\|\leqslant 1/(1-\varepsilon)$. Hence the infimum in (5.6.3) is at most $1/(1-\varepsilon)$; thus at most 1. On the other hand, the infimum is at least 1, for $\|y\|\geqslant|\eta(y)|\geqslant 1$. This proves the formula for DT_Y in the complement of Y. We note that for open half-spaces the formula is the same. For $x\in Y$ we can therefore use the symmetry $\mathsf{DT}_{\complement Y} = -\mathsf{DT}_Y$.

We note that if Y is a half-space which is not closed (equivalently, defined by a discontinuous linear functional), then $\mathsf{DT}_Y=0$ identically.

Theorem 5.6.2. *Let A be any convex set in a normed vector space E. Then its distance transform is*

$$\mathsf{DT}_A = \sup_Y \left(\mathsf{DT}_Y; Y \textit{ is a closed half-space containing } A\right).$$

Proof. We first note that the theorem is trivially true if $A=\emptyset$ or $A=E$. Assume now that $A\neq\emptyset, E$.

For every set Y containing A we have $\mathsf{DT}_Y\leqslant\mathsf{DT}_A$, so the supremum in the statement of the theorem can never exceed DT_A.

If $x\notin\overline{A}$ we consider the open ball $B_<(x,r)$, where $r=\mathsf{DT}_A(x)>0$. This ball and A are disjoint. By the Hahn–Banach theorem there is a hyperplane separating the two. This means that there is a continuous linear form ξ such that $\xi(y)\leqslant\xi(z)$ for all $y\in A$ and all $z\in B_<(x,r)$. Thus the closed half-space Y defined by $\xi(y)\leqslant \inf_{z\in B_<(x,r)}\xi(z)$ contains A, and $\mathsf{DT}_Y(x)\geqslant r=\mathsf{DT}_A(x)$. This shows that the supremum in the statement of the theorem attains DT_A at the point x.

Now take $x\in A$. Then there is an open ball $B_<(x,r)$ with maximal radius $r=-\mathsf{DT}_A(x)\geqslant 0$ contained in A. (Note that $r=0$ is allowed this time.) Take $s=r+\varepsilon>r$, where ε is arbitrarily small. There exists a point b in $B_{\leqslant}(x,s)\smallsetminus A$. By the Hahn–Banach theorem again there is a closed hyperplane passing through b with A on one side. The corresponding half-space Y solves our problem, for $-\mathsf{DT}_Y(x)$ cannot be larger than $d(x,b)\leqslant s$, so that $\mathsf{DT}_Y(x)\geqslant -s=\mathsf{DT}_A(x)-\varepsilon$. Since ε is arbitrary, this proves the equality in A, and by continuity also in its closure $\overline{A}$. We already proved the equality in the complement of $\overline{A}$, so we are done.

Corollary 5.6.3. *The distance transform of an arbitrary convex subset of a normed vector space is a convex function.*

Proof. The distance transform of a half-space is an affine function as we have seen. Every affine function is convex, and the supremum of any family of convex functions is convex.

Definition 5.6.4. *Given any subset A of a vector space E we define its* supporting function H_A *by*

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

Here $E^\star$ is the space of all linear forms on E. It is called the *algebraic dual* of E, and contains the dual E', maybe strictly.

Example 5.6.5. If A is a ball, $A = B_\leqslant(c, r)$ with $r \geqslant 0$, or $B_<(c, r)$ with $r > 0$, then $H_A(\xi) = \xi(c) + r\|\xi\|'$, $\xi \in E^\star$, where $\|\cdot\|$ is an arbitrary norm and $\|\cdot\|'$ its dual norm. Here we interpret the product $r\|\xi\|'$ as 0 if $r = 0$.

Corollary 5.6.6. *The distance transform of a closed convex subset A of a normed space E is equal to*

$$\mathsf{DT}_A(x) = \sup_{\|\eta\|'=1} \big(\eta(x) - H_A(\eta)\big), \qquad x \in E.$$

Proof. We know that the distance transform at a point x is equal to the supremum of all values $\mathsf{DT}_Y(x)$ when Y varies in the family of all closed half-spaces containing A. However, every such half-space is contained in a minimal half-space, and the minimal half-spaces $Y = \{y; \eta(y) \leqslant \alpha\}$ are precisely those for which $\alpha = H_A(\eta)$, thus with distance transform $\eta(x) - H_A(\eta)$.

Definition 5.6.7. *To any function $\varphi\colon E \to [-\infty, +\infty]$ we define its* Fenchel[15] transform *by $\widetilde{\varphi}\colon E^\star \to [-\infty, +\infty]$ on the algebraic dual $E^\star$ of E by*

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

The Fenchel transform generalizes the supporting function: $\widetilde{i_A} = H_A$.

Example 5.6.8. If the epigraph of a function φ is a paraboloid, $\varphi(x) = a + \beta\cdot x + \frac{1}{2}c\|x\|_2^2$, $x \in \mathbf{R}^n$, where $a \in \mathbf{R}$, $\beta \in \mathbf{R}^n$ and $c > 0$, then the same is true of the epigraph of its transform: $\widetilde{\varphi}(\xi) = -a + \frac{1}{2}c^{-1}\|\xi - \beta\|_2^2$.

Corollary 5.6.6 shows that the distance transform is the Fenchel transform of the supporting function restricted to the unit sphere—more precisely, if we define a function g to be equal to H_A on the unit sphere S' and $+\infty$ elsewhere, then $\mathsf{DT}_A = \widetilde{g}$. We may therefore ask what happens if we apply the transformation again. To this end we need a general result on the iterated Fenchel transform. We let F be any linear subspace of $E^\star$ and define the Fenchel transform of any function f defined on F by

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

[15] Named for Werner Fenchel, 1905—1988.

In particular we may form the second transform $\tilde{\tilde{\varphi}}$ of a function defined on E. The main result in the theory of the Fenchel transformation is this:

Theorem 5.6.9. *Let φ be a function defined on a vector space E and let F be any subspace of its algebraic dual $E^\star$. Then we always have $\tilde{\tilde{\varphi}} \leqslant \varphi$. Equality holds if and only if*
(A) *φ is convex;*
(B) *φ is lower semicontinuous for the weakest topology for which all linear forms in F are continuous; and*
(C) *φ does not take the value $-\infty$ unless it is identically equal to $-\infty$.*

Property (B) here means that if $\varphi(x) > a$, then there are linear forms $\xi_1, \dots, \xi_m \in F$ such that $\varphi(y) > a$ when $|\xi_j(y-x)| \leqslant 1$, $j = 1, \dots, m$. In $\mathbf{R}^n$ we usually choose $F = \mathbf{R}^n$; the semicontinuity is then semicontinuity with respect to the usual topology of $\mathbf{R}^n$. The necessity of the three properties is easy to prove; the sufficiency will be accepted here without proof.

Corollary 5.6.10. *If A is a nonempty closed convex subset of E, the Fenchel transform $\tilde{\varphi}$ of its distance transform $\varphi = \mathsf{DT}_A$ is*

$$\tilde{\varphi}(\xi) = \begin{cases} H_A(\xi), & \|\xi\|' = 1; \\ +\infty, & \|\xi\|' > 1; \\ \textit{the largest convex minorant of } H_A\big|_{S'}, & \|\xi\|' < 1. \end{cases}$$

Here S' is the unit sphere for the dual norm $\|\cdot\|'$. If A is empty, then $\tilde{\varphi}$ is $-\infty$ identically.

Proof. Let g be equal to H_A on the unit sphere S' and $+\infty$ elsewhere, and h equal to H_A on the closed unit ball of E' and equal to $+\infty$ elsewhere. As already noted, DT_A is the Fenchel transform of g, $\mathsf{DT}_A = \tilde{g}$. The transform of $\varphi = \mathsf{DT}_A$ is therefore the second Fenchel transform of g: $\tilde{\varphi} = \tilde{\tilde{g}}$. Theorem 5.6.9 tells us that $\tilde{\tilde{g}}$ is the largest lower semicontinuous convex minorant of g. Since $h \leqslant g$ we have $\tilde{\tilde{h}} \leqslant \tilde{\tilde{g}}$, and we know that $\tilde{\tilde{h}} = h$ since h is convex, lower semicontinuous, and never takes the value $-\infty$; $h = \tilde{\tilde{h}} \leqslant \tilde{\tilde{g}} \leqslant g$. Therefore, since h and g agree on S', $\tilde{\tilde{g}} = g = H_A$ on S'. Outside B' (in particular in $E^\star \smallsetminus E'$) we have $g = h = +\infty$, so $\tilde{\varphi} = \tilde{\tilde{g}} \geqslant \tilde{\tilde{h}} = +\infty$ there. Finally we note that, since g is plus infinity in the open unit ball, $\tilde{\tilde{g}}$ is as described in the statement of the corollary in that ball.

All this holds if A is nonempty; otherwise we easily see that $\tilde{\varphi} = H_A = -\infty$ identically.

5.7. Exercises

5.1. (a) Prove that, with the notation of Theorem 5.1.4, we have $\alpha(A) = 1$ and $\beta(A) = 3$ when $X = \mathbf{Z}^n$ with the l^∞ metric and $A \neq \emptyset, \mathbf{Z}^n$.

(b) Determine the quantities $\alpha(A)$ and $\beta(A)$ when $X = \mathbf{Z}^n$ with the l^1 metric.

5.2. Prove Proposition 5.2.2.

5.3. Let A be the (filled) square $\{x \in \mathbf{R}^2; |x_1|, |x_2| \leqslant 1\}$ and B a Euclidean closed disk of radius r, viz. $B = \{x \in \mathbf{R}^2; x_1^2 + x_2^2 \leqslant r^2\}$. Describe for all positive r the seven sets $A + B$, $A \ominus B$, $B \ominus A$, A^B, B^A, A_B, B_A. Calculate the distance transforms DT_A and DT_B when the distance is given by $d(x, y) = \|x - y\|_p$, $p = 1, 2, \infty$.

5.4. (a) Given a subset A of an abelian group with a translation-invariant metric, show that $\mathsf{DT}_{A+B_<(0,r)} = \mathsf{DT}_A - r$ holds in $\complement(A + B_<(0, r))$ (cf. Proposition 5.1.6). (b) Show that it does hold everywhere if the space is a vector space and A is convex. (c) What about $\mathbf{Z}^2$?

5.5. Let X be a metric space with metric d. Define a metric d_α in $X \times \mathbf{R}$ by

$$d_\alpha\big((x, s), (y, t)\big) = \alpha d(x, y) + |s - t|, \qquad (x, s), (y, t) \in X \times \mathbf{R},$$

where α is any positive number. Prove that, for any subset A of X,

$$\mathsf{DT}_{\operatorname{epi} \mathsf{DT}_A}(x, s) = \mathsf{DT}_A(x) - s, \qquad (x, s) \in X \times \mathbf{R},$$

if $\alpha \geqslant 2$ ($\alpha \geqslant 1$ in a vector space). Show by examples that this is not necessarily so if $0 < \alpha < 2$ ($0 < \alpha < 1$ in a vector space).

5.6. Square pixels. Assume that the centers of square pixels are placed at the points $\mathbf{Z}^2$ with integer coordinates in $\mathbf{R}^2$. Since $\mathbf{Z}^2$ is group, we can use infimal convolution. Actually we can work indifferently in $\mathbf{R}^2$ or $\mathbf{Z}^2$. Define a function $F \colon \mathbf{R}^2 \to [0, +\infty]$ by $F(0) = 0$, $F(\pm 1, 0) = F(0, \pm 1) = a > 0$, $F(\pm 1, \pm 1) = b > 0$ (four points with value b), and $F(x) = +\infty$ at all other points. Then let $f = \lim F \mathbin{\square} F \mathbin{\square} \cdots \mathbin{\square} F$ as the number of terms tends to infinity.

(a) Show that if $a = 1$, $b = \sqrt{2}$, then

$$1 \leqslant \frac{f(x)}{\|x\|_2} \leqslant \sqrt{\frac{2\sqrt{2}}{\sqrt{2} + 1}} \approx 1.08239, \qquad x \in \mathbf{Z}^2.$$

(b) Show that the best choice for a and b if we want to have

$$1 \leqslant \frac{f(x)}{\|x\|_2} \leqslant C, \qquad x \in \mathbf{Z}^2,$$

with C as small as possible is $a = 1$, $b = \sqrt{2}$.

(c) For all $a, b > 0$ there is a constant C depending on a, b such that

$$C^{-1} \leqslant \frac{f(x)}{\|x\|_2} \leqslant C, \qquad x \in \mathbf{Z}^2.$$

Take $b = a\sqrt{2}$ and find the a which renders C as small as possible. Show that the smallest possible value of C is

$$C = \Lambda(f, \|\cdot\|_2) = \sqrt[4]{\frac{4}{2 + \sqrt{2}}} \approx 1.04038,$$

and that it is attained when a and $b = a\sqrt{2}$ take the values (5.4.3).

(d) Now vary both a and b. Show that the smallest C such that $C^{-1} \leqslant f(x)/\|x\|_2 \leqslant C$ for all $x \in \mathbf{Z}^2$ is the same as in (c), i.e., that we do not gain anything by taking b different from $b = a\sqrt{2}$.

(e) Prove that if we use $\Lambda'(f, \|\cdot\|_2)$ to measure deviation from the Euclidean distance, then the optimal values are as indicated in (5.4.4).

(f) Prove that if we use $\Lambda'(\|\cdot\|_2, f)$ to measure deviation from the Euclidean distance, then the optimal values are as indicated in (5.4.5).

5.7. Hexagonal pixels. The centers of hexagonal pixels form a triangular pattern: if we identify $\mathbf{R}^2$ with $\mathbf{C}$ for convenience they can be placed at the points $p + q\omega$, $p, q \in \mathbf{Z}$, where $\omega = \frac{1}{2} + i\frac{\sqrt{3}}{2}$.

(a) Define a function $F\colon \mathbf{C} \to [0, +\infty]$ by $F(0) = 0$, $F(\pm 1) = F(\pm\omega) = F(\pm(1-\omega)) = 1$, and $F(x) = +\infty$ at all other points. Then let $f = \lim F \,\square\, F \,\square\, \cdots \,\square\, F$ as the number of terms tends to infinity. Show that

$$1 \leqslant \frac{f(x)}{\|x\|_2} \leqslant \frac{2}{\sqrt{3}} \approx 1.15470, \qquad x \in \mathbf{Z} + \mathbf{Z}\omega.$$

Thus a suitable multiple f_1 of f has $\Lambda(f_1, \|\cdot\|_2) \leqslant \sqrt{2}/\sqrt[4]{3} \approx 1.07457$.

(b) Now define $F\colon \mathbf{Z}^2 \to [0, +\infty]$ by $F(0) = 0$, $F(\pm 1, 0) = F(0, \pm 1) = F(1, -1) = F(-1, 1) = 1$ and $F(x) = +\infty$ at all other points. Let $f = \lim F \,\square\, F \,\square\, \cdots \,\square\, F$ as the number of terms tends to infinity. Find a Euclidean norm $\|\cdot\|$, i.e., a norm defined by an inner product, such that

$$\|x\| \leqslant f(x) \leqslant C\|x\|, \qquad x \in \mathbf{Z}^2,$$

with C as small as possible.

5.8. Triangular pixels. The centers of triangular pixels form a hexagonal pattern. These hexagonally placed points do not form a subgroup of $\mathbf{R}^2$, and therefore infimal convolution cannot be applied directly as in the case of square or hexagonal pixels. However, if we take two steps in the hexagons, we get a group of triangularly placed points, which is isometric to the group $\mathbf{Z} + \mathbf{Z}\omega$ formed by the hexagonal pixels considered in exercise 5.7. For an even number of steps, the inequalities for hexagonal pixels can be applied; for an odd number of steps, we first take an even number of steps and then one extra step. Investigate what the inequalities of exercise 5.7 yield.

5.9. Prove that in an abelian group with a translation-invariant metric which is upper regular for the triangle inequality, $B_{\leqslant}(a, r) = B_{\leqslant}(b, s)$ implies that $a = b$.

5.10. Prove that the chamfer distance in $\mathbf{Z}^2$ defined by $F(\pm 1, 0) = F(0, \pm 1) = 3$, $F(\pm 1, \pm 1) = 4$ is not lower regular for the triangle inequality.

5.11. Calculate the values of the functions ρ, σ, ρ_-, and σ_+ when the space is $\mathbf{Z}^2$ and the distance is given by the l^2-norm, say for real $r \leqslant 4$.

6. Skeletonizing

6.1. Definition of the skeleton

If A is any subset of a metric space X, then its interior A° is the union of all open balls contained in A. This is typically the union of a very large family of sets. We would like to describe A° as the union of a smaller family. It is obvious that if we have two balls contained in A, $B_<(a, r)$ and $B_<(b, s)$, and one is contained in the other, then we may throw away the smaller ball without changing the union. In fact, for every ball $B_<(a, r)$ in the union, we may throw away all balls contained in that ball without changing the union. This leads to the concept of a maximal ball. A maximal ball must be retained, but all balls contained in a maximal ball may be dispensed with.

The importance of skeletons in applications is due to the fact that they are thin in some sense but nevertheless retain important information about an object, for instance its general shape, and that, given the skeleton and the distance transform at the points in the skeleton, we can reconstruct the whole object. Typically we save memory when listing only the skeleton and the quench function.

If a is the center of a maximal open ball $B_<(a,r)$ contained in a set A, then necessarily $r = -\mathsf{DT}_A(a)$. In fact, when we defined the distance transform $\mathsf{DT}_A(a)$ at a point a, we looked at all balls with center a contained in A and we took the largest such ball. Note that then we kept the center fixed. There is a largest ball with center a, which in particular is maximal among these balls. By way of contrast, when we define the skeleton we shall vary both the center and the radius and look at all balls contained in A, regardless of their centers. We shall now give a name to the centers of maximal balls.

Definition 6.1.1. *Let A be a subset of a metric space X. We define the* skeleton[16] *of A, denoted by $\mathsf{Sk}(A)$, as the set of all centers of maximal nonempty open balls contained in A.*

The definition means that $a \in \mathsf{Sk}(A)$ if and only if there exists a number $r > 0$ such that $B_<(a,r) \subset A$ and such that if a ball $B_<(b,s)$ is contained in A and contains $B_<(a,r)$, then $B_<(b,s) = B_<(a,r)$.[17] The skeleton may be empty: think of a set with empty interior or of a half-space in $\mathbf{R}^n$. A half-space contains lots of balls, but there are no maximal balls. So obviously we need to investigate whether there exist maximal balls—and whether there are enough of them in the formation of the interior of A. For this we shall need Zorn's Lemma.

6.2. Existence of skeletons

An ordered set X is said to be *totally ordered* if for any two elements $x, y \in X$ we have $x \leqslant y$ or $y \leqslant x$. An ordered set X is said to be *inductive* or *inductively ordered* (Bourbaki 1963:34) if every totally ordered subset of X possesses a majorant in X.[18] This means that for every $Y \subset X$ which is totally ordered, there exists an element $b \in X$ such that $y \leqslant b$ for all $y \in Y$. This concept is of interest because it is used as an hypothesis in Zorn's Lemma, which guarantees the existence of maximal elements.

Theorem 6.2.1 (Zorn's Lemma). *Every inductively ordered set possesses a maximal element.*

We shall accept Zorn's Lemma here. One can prove it using the Axiom of Choice; conversely, the latter can be proved from Zorn's Lemma. To establish the existence of a maximal element we shall have to prove that a certain order is inductive.

Theorem 6.2.2. *Let $\mathbf{Z}^n$ be equipped with a metric which either is inherited from a norm on $\mathbf{R}^n$ or a chamfer distance in the sense of Definition 5.3.7, and let A be a finite subset. Then the set of all open balls contained in A is inductively ordered.*

Proof. Let us consider a union $A_M = \bigcup_{(c,r)\in M} B_<(c,r)$ of a family of open balls contained in A, where M is a subset of $\mathbf{Z}^n \times \mathbf{R}$. Assume that the family is totally

[16] It seems that the term *skeleton* was first used in this context by Rosenfeld & Pfaltz (1966).

[17] I did not say that $b = a$ and $s = r$.

[18] The empty set is totally ordered, so an inductively ordered set must be nonempty.

ordered, i.e., for any two pairs $(a, r), (b, s) \in M$, either $B_<(a, r)$ is a subset of $B_<(b, s)$ or conversely. Clearly A_M, being a subset of A, is finite, which implies that it is equal to one of the balls $B_<(c, r)$ with $(c, r) \in M$. We are done.

In $\mathbf{R}^n$ things are less simple.

Theorem 6.2.3. *Let A be a set in a finite-dimensional normed vector space E. Assume that A is bounded. Then the set of open balls contained in A is inductively ordered.*

Proof. Let as before $A_M = \bigcup_{(c,r)\in M} B_<(c, r)$ be a union of open balls contained in A, where M is a subset of $E \times \mathbf{R}$. We assume that the family of these balls is totally ordered. Define R as the supremum of all numbers r such that there exists a $c \in E$ such that $(c, r) \in M$. Since A is bounded, this supremum is necessarily finite. For every $j = 1, 2, 3, \ \dots$ there exists a number $r_j > R - 1/j$ and a point c_j in A such that $(c_j, r_j) \in M$. Unless one of the radii is equal to R we may also choose the r_j so that $r_{j+1} > r_j$. Since the sequence of centers (c_j) is bounded, it has a converging subsequence; let us change notation so that (c_j) itself is converging. Let its limit be C. We claim that A_M is equal to the ball $B_<(C, R)$.

We shall prove first that A_M is contained in $B_<(C, R)$. Let $x \in A_M$; there exists $(c, r) \in M$ such that $d(x, c) < r$. Define $\varepsilon = r - d(x, c) > 0$. Then the ball $B_<(x, \varepsilon)$ is contained in $B_<(c, r)$. If $r = R$, then we must have $B_<(c, r) = B_<(C, R)$, in particular x belongs to $B_<(C, R)$. If, on the other hand $r < R$, then we take k so large that $r_k > r$, and we must have $B_<(c, r) \subset B_<(c_k, r_k)$ since the opposite inclusion is impossible and one of them must hold by hypothesis. Thus $B_<(x, \varepsilon) \subset B_<(c_j, r_j)$ for all $j \geqslant k$, which implies, on letting j tend to infinity, that $B_<(x, \varepsilon) \subset B_\leqslant(C, R)$. But then x belongs also to the open ball $B_<(C, R)$ and we are done. (Note that this part of the proof is not valid in a general metric space.)

In the other direction, let us prove that $B_<(C, R)$ is contained in A_M. Take any point $x \in B_<(C, R)$. Then $d(x, C) < R$, and we may define $\varepsilon = R - d(x, C) > 0$. We then have $d(x, c_j) \leqslant d(x, C) + d(C, c_j) = R - \varepsilon + d(C, c_j)$, where the right-hand side is less than r_j for large j. Hence x belongs to $B_<(c_j, r_j)$, and therefore to A_M. (This part of the proof is valid in any compact metric space.)

If the norm is Euclidean, it is enough to assume that A does not contain a half-space. Also, for any given norm in $\mathbf{R}^n$, it is enough to assume that A does not contain a cone of a certain aperture.

Example 6.2.4. One might think that the result should hold in any compact metric space. However, simple examples show that this is not so. Define a compact metric space X as consisting of the segment $A = [-1, 1] \times \{0\}$ in $\mathbf{R}^2$ and adjoin to it the point $(0, 1)$. The metric shall be that induced by the usual Euclidean metric in $\mathbf{R}^2$. Consider now the open balls $B_<(c_j, r_j)$, where $c_j = (1/j, 0)$ and $r_j = 1$, $j = 1, 2, \ \dots$. Note that $B_<(c_j, r_j) = \left]-1 + 1/j, 1\right] \subset A$. The sequence of balls is increasing with j, and its union is the segment $A_M = \left]-1, 1\right] \times \{0\}$. However, this segment is not an open ball in X. What is worse: it is not contained in an open ball in A. Hence the family of open balls in A is not inductively ordered.

There is a well-known trick to get rid of the hypothesis that A must not contain any half-space. Indeed a half-space is a limiting case of a ball, and if we compactify $\mathbf{R}^n$ by adding a point at infinity, then $\mathbf{R}^n \cup \{\infty\}$ can be regarded as the n-sphere. The balls in $\mathbf{R}^n$ then become spherical caps on the sphere, and so do the half-spaces. But these spherical caps are actually balls for a suitable metric on the sphere. So $\mathbf{R}^n \cup \{\infty\}$ is a metric space and the open balls contained in any given subset are inductively ordered—the proof is very much like the one we have done in $\mathbf{R}^n$. The compactification allows us to define a generalized skeleton, which may contain ∞, but we need of course to modify the quench function.

Corollary 6.2.5. *Let A be a bounded subset of a finite-dimensional normed vector space, or a bounded subset of $\mathbf{Z}^n$, where $\mathbf{Z}^n$ is provided with a metric as in Theorem 6.2.2. The union of all open balls with center c belonging to the skeleton and radius equal to $-\mathsf{DT}_A(c)$ is equal to the interior of A. In particular, if A has interior points, then the skeleton of A is nonempty.*

Proof. Take any point $x \in A^\circ$. The ball $B_<(x, \varepsilon)$ is contained in A° for some small positive ε. By Zorn's lemma (Theorem 6.2.1) and Theorem 6.2.2 or 6.2.3, respectively, there is a maximal ball $B_<(c, r)$ containing $B_<(x, \varepsilon)$ and contained in A. Thus $c \in \mathsf{Sk}(A)$ and $x \in B_<(c, r)$, with $r = -\mathsf{DT}_A(c)$.

In any metric space where the conclusion of Theorem 6.2.3 holds we have $A^\circ = \bigcup_{c \in \mathsf{Sk}(A)} B_<(c, -\mathsf{DT}_A(c))$. Here $-\mathsf{DT}_A(c) = (\mathsf{DT}_A(c))^-$ is the quench function evaluated at c. Knowledge of $\mathsf{Sk}(A)$ and the restricion of DT_A to $\mathsf{Sk}(A)$ is equivalent to knowing A°. This shows how we can reconstruct A° from $\mathsf{Sk}(A)$ and the quench function. However, it is sometimes not necessary to use even all the points in the skeleton, e.g., when A is the union of two disks; see example 6.3.1 below.

6.3. Properties of skeletons

In some sense the skeleton is a thin set. For instance, it is easy to prove that a skeleton in $\mathbf{R}^n$ has no interior points (cf. exercises 6.1 and 6.2). On the other hand, the closure of the skeleton need not be of Lebesgue measure zero (cf. exercise 6.3). These results are mentioned by Serra (1982:378) and Matheron (1988:218). It is also stated there that it is unknown whether the skeleton has Lebesgue measure zero, and whether the interior of its closure is empty.[19]

The skeleton has, generally speaking, bad continuity properties.

Example 6.3.1. Let D be the open unit disk in $\mathbf{R}^2$, $D = \{(x, y) \in \mathbf{R}^2; x^2 + y^2 < 1\}$. Its skeleton is just the origin. Then add a small open disk D_ε with center at $(1, 0)$ and radius $\varepsilon > 0$. The skeleton of the new set $A = D \cup D_\varepsilon$ is the entire segment $[0, 1] \times \{0\}$ for all small positive ε. Thus a very small change in the set causes the skeleton to grow. Note that here it is not necessary to use all the points in the skeleton to reconstruct A: it suffices to take the disks with centers at 0 and $(1, 0)$. Even more dramatic is perhaps the growth in the skeleton when we remove a small closed disk: consider $D \smallsetminus \overline{D_\varepsilon}$.

In $\mathbf{Z}^2$ the continuity properties are of course different, but a small change can still cause points to appear far from the original skeleton.

Example 6.3.2. Let $A = [-m, m]_{\mathbf{Z}} \times [-m, m]_{\mathbf{Z}}$ be a large square in $\mathbf{Z}^2$. Its skeleton for the chess-board metric is just the origin. If we add a single point $(m + 1, 0)$ to A, the skeleton of the new set is $\{0, (m + 1, 0)\}$. What happens if we remove a point? Consider $A \smallsetminus \{(m, 0)\}$.

[19] It seems that the answers to these questions are still unknown.

The skeleton of a set A in $\mathbf{R}^2$ need not be a closed set, even if A has a smooth boundary.

Example 6.3.3. Let first $U = A + B$, the dilation of the segment $A = [(-1,1),(1,1)]$ by the unit disk $B = \{(x,y) \in \mathbf{R}^2; x^2 + y^2 < 1\}$. This is an open set and its skeleton is $\mathsf{Sk}(U) = A$. Then modify U as follows. The lower boundary of U is just the segment $[(-1,0),(1,0)]$, so that U is locally defined by the inequality $y > 0$ near this segment. We replace it by an inequality $y > \varphi(x)$, where $\varphi(x) = \sum_1^\infty c_j \psi(2^j x - 1)$, ψ being an even, nonpositive function in C_0^∞ with $\psi(0) = -1$ and support contained in $[-\frac{1}{3}, \frac{1}{3}]$. For a suitable choice of positive constants c_j, the function φ is in C^∞, and the skeleton of the new open set V so defined contains segments $\{2^{-j}\} \times [a,b]$ for $j = 1, 2, 3, \ldots$ for a suitable choice of a and $b > a$. But no point on the limiting segment $\{0\} \times [a,b]$, belongs to it. (This example is essentially taken from (Matheron 1988:219).)

We shall now give a characterization of points in the skeleton. The following result was proved in $\mathbf{R}^n$ by Matheron (1988:225).

Theorem 6.3.4. *Let E be a normed space with metric given by the norm: $d(x,y) = \|x - y\|$. Let A be a nonempty proper subset of E, fix a point c in the interior of A, and define $h(x) = d(x,c) + \mathsf{DT}_A(x)$, $x \in E$. Then c belongs to the skeleton of A if and only if h has a minimum only at c.*

Proof. If $B_<(c,r) \subset B_<(x,s) \subset A$, where $r = -\mathsf{DT}_A(c) > 0$, then $s \geqslant d(c,x) + r$. This implies that $h(x) = d(x,c) + \mathsf{DT}_A(x) \leqslant d(x,c) - s \leqslant d(c,c) - r = h(c)$. If c is the only point where h attains its minumum, we must have $x = c$ and it follows that $B_<(c,r)$ is maximal, hence that $c \in \mathsf{Sk}(A)$.

Conversely, assume that c is in the skeleton and that x is a point where $h(x) \leqslant h(c)$. Then $d(c,x) - s \leqslant -r$, where we define $r = -\mathsf{DT}_A(c)$, $s = -\mathsf{DT}_A(x)$. This implies that $B_<(x,s)$ contains $B_<(c,r)$. Since c is in the skeleton of A, the two balls must be equal, which implies that $x = c$. Therefore the infimum of h is attained at c and only there.

Thanks to the calculus of balls developed in section 5.5 we can generalize this result to other groups. In a normed space of positive dimension, the open ball of radius $r = -\mathsf{DT}_A(a)$ is the interior of the closed ball of the same radius and the same center. In a group where the set of distances is discrete, the open ball $B_<(a,r)$ can be described as the closed ball of radius $\rho_-(r)$. Since the conditions for working with closed balls are more easily satisfied than those for open balls, we will get a more applicable result if we replace the function $x \mapsto d(x,c) + \mathsf{DT}_A(x)$ by $x \mapsto d(x,c) - \rho_-(-\mathsf{DT}_A(x))$.

Theorem 6.3.5. *Let G be an abelian group with a translation-invariant metric d which is upper regular for the triangle inequality and such that the set of all distances is discrete. Let A be a nonempty proper subset of G, fix a point $c \in A$, and define $h(x) = d(x,c) - \rho_-(-\mathsf{DT}_A(x))$, $x \in G$. Then c belongs to the skeleton of A if and only if h has a minimum only at c.*

Proof. If $B_<(c,r) \subset B_<(x,s) \subset A$, where $r = -\mathsf{DT}_A(c)$, then by Theorem 5.2.4 $\rho_-(s) \geqslant d(c,x) + \rho_-(r)$. This implies that $h(x) = d(x,c) - \rho_-(-\mathsf{DT}_A(x)) \leqslant d(x,c) - \rho_-(s) \leqslant d(c,c) - \rho_-(r) = h(c)$. If c is the only point where h attains its minumum, we must have $x = c$ and it follows that $B_<(c,r)$ is maximal, hence that $c \in \mathsf{Sk}(A)$.

Conversely, assume that c is in the skeleton and that x is a point where $h(x) \leqslant h(c)$.

Then $d(c,x) - \rho_-(s) \leqslant -\rho_-(r)$, where we define $r = -\mathsf{DT}_A(c)$, $s = -\mathsf{DT}_A(x)$. This implies that $B_\leqslant(x, \rho_-(s))$ contains $B_\leqslant(c, \rho_-(r))$. But $B_\leqslant(x, \rho_-(s)) = B_<(c,s)$ and $B_\leqslant(c, \rho_-(r)) = B_<(c,r)$ by Proposition 5.5.3. Since c is in the skeleton of A, the two balls must be equal, which implies that the difference between their centers belongs to a bounded subgroup, thus that $x = c$. Therefore the infimum of h is attained at c and only there.

6.4. Exercises

6.1. Show that in $\mathbf{R}^n$ with a Euclidean metric every skeleton has an empty interior.

6.2. Let G be an abelian group and P any nonempty subset of G. A set $A \subset G$ will be called *P-open* if it is if the form $\bigcup_{x \in M}(x + P)$ for some $M \subset G$. The set of all P-open sets form what we might call a semitopology τ_P on G: an arbitrary union of P-open sets is P-open. Assume now that P is a finite symmetric subset of G, and let F and f be as in Corollary 5.3.6 with F symmetric. Let finally $\mathsf{Sk}(A)$ be the skeleton of a set A defined by the chamfer distance $d(x,y) = f(x-y)$. Prove that interior of $\mathsf{Sk}(A)$ defined by τ_P is empty: there is no nonempty P-open set contained in $\mathsf{Sk}(A)$.

6.3. Construct an example of an open set in $\mathbf{R}^2$ such that the closure of its skeleton with respect to a Euclidean metric is of positive Lebesgue measure. *Hint:* Take U as the unit disk, and add to it denumerably many equilateral triangles with two points on the circumference and the third point with rational argument. This will yield a set whose skeleton contains a lot of rays; the limits of these rays have arguments in a set of measure $2\pi - \varepsilon$.

6.4. Show that if we use the l^∞ distance there exists a bounded open connected set in $\mathbf{R}^2$ such that its skeleton is not connected.

6.5. We define the *r-skeleton*, associated to the radius $r > 0$, as the set $\mathsf{Sk}_r(A)$ of points in the skeleton where the quench function is equal to r. Therefore $\mathsf{Sk}(A) = \bigcup_{r>0} \mathsf{Sk}_r(A)$.

Assume now that we are in a normed vector space with distance $d(x,y) = \|x - y\|$.

(a) Show that the reconstruction of the interior of a bounded set A takes the form

$$A^\circ = \bigcup_{r>0} \big(\mathsf{Sk}_r(A) + B_<(0,r)\big).$$

(b) Show that we can also reconstruct the dilations and erosions by balls from the r-skeletons for bounded open sets A:

$$A + B_<(0,r) = \bigcup_{s>0} \big(\mathsf{Sk}_s(A) + B_<(0, r+s)\big), \qquad r > 0;$$

$$A \ominus B_\leqslant(0,r) = \bigcup_{s>r} \big(\mathsf{Sk}_s(A) + B_<(0, s-r)\big), \qquad r > 0.$$

(c) Find some conditions on a set A which guarantee that the skeleton of its dilation by a ball is the same as that of A:

$$\mathsf{Sk}\big(A + B_<(0,r)\big) = \mathsf{Sk}(A), \qquad r > 0.$$

(d) Find conditions on a set A which ensure that the skeleton of an erosion is simply obtained by deleting a certain part of the skeleton:

$$\mathsf{Sk}(A \ominus B_\leqslant(0,r)) = \bigcup_{s>r} \mathsf{Sk}_s(A), \qquad r > 0.$$

(e) Show that

$$\mathsf{Sk}_r(A) = \bigcap_{s>0} \Big[\big(A \ominus B_<(0,r)\big) \smallsetminus \big(A \ominus B_<(0,r)\big)_{B_\leqslant(0,s)} \Big], \qquad r > 0.$$

6.6. What remains of exercise 6.5 if we only know that we are in an abelian group with a translation-invariant metric?

7. Lattices

7.1. Definition and first properties of lattices

Lattice theory was developed by Birkhoff[20] and others in the beginning of the twentieth century (Birkhoff 1940, 1948). There is an analogy between lattice theory and the theory of vector spaces. The theory of topological vector spaces was developed to a large extent because of the theory of distributions, which in turn was motivated by applications in partial differential equations. Developments in image processing motivated a renewed interest in lattice theory, in particular in complete lattices. Lattice theory was applied to switching circuits, and it was then enough, because of general finiteness conditions, to form models using lattices, but in image processing it is more convenient to assume completeness.

While vector spaces are useful in modelling linear problems, lattices seem to be more adapted to nonlinear problems. Auditory phenomena are often additive: all the instruments of an orchestra can be heard; while with visual phenomena this is not so: one object can block another from our view. This indicates that linear models may suffice for the first kind of phenomena, while the visual ones are more in agreement with nonlinear operators like supremum and infimum.

As we shall see, there are also analogies between preordered sets, in particular lattices, and topological spaces. The increasing mappings in the first case correspond to continuous linear mappings in the second.

Let L be an ordered set and A a subset of L. An element $b \in L$ is said to be the *infimum* of all elements $a \in A$ if b is the largest minorant of all $a \in A$. This means that $b \leqslant a$ for all elements $a \in A$, and that if $b' \leqslant a$ for all $a \in A$, then $b' \leqslant b$. The infimum, if it exists, is necessarily unique. The infimum of the empty set exists if and only if L possesses a largest element, and if so, the infimum is this largest element.

We shall write

$$b = \inf_{a \in A} a = \inf(a; a \in A) = \bigwedge_{a \in A} a$$

for the infimum of all elements in A; if A has only n elements we write $b = a_1 \wedge \cdots \wedge a_n$, in particular $b = a_1 \wedge a_2$ if $n = 2$.

Similarly we define the *supremum*

$$c = \sup_{a \in A} a = \sup(a; a \in A) = \bigvee_{a \in A} a$$

as the smallest majorant of all elements in A. The supremum of the empty set, $\sup_{x \in \varnothing} x$, exists if and only if L has a smallest element.

[20] Garrett Birkhoff, 1911—1996.

If any set consisting of two elements in L has an infimum, we shall call L an *inf-semilattice*; similarly, if any two-set of L has a supremum, we shall call L a *sup-semilattice.* If L is both an inf-semilattice and a sup-semilattice we shall call L a *lattice.*

If any nonempty subset, finite or infinite, has an infimum, L will be said to be a *complete inf-semilattice*; analogously we define *complete sup-semilattice* and *complete lattice.* A complete inf-semilattice has a smallest element, which is the infimum of all elements, denoted by $\mathbf{0}$, and a complete sup-semilattice has a largest element, the supremum of all elements, denoted by $\mathbf{1}$.

A complete inf-semilattice with a largest element $\mathbf{1}$ is also a complete lattice. Indeed, the supremum of any set of elements is equal to the infimum of all majorants of the set—this set is not empty since $\mathbf{1}$ is a majorant.

In a complete lattice, the infimum of the empty set exists and is $\mathbf{1}$, and the supremum of the empty set is $\sup_{x\in\varnothing} x = \mathbf{0}$.

It is possible to define a lattice as a set with two binary operations $\wedge$ and $\vee$ satisfying certain axioms.

A *sublattice* is defined just like a subgroup with respect to the operations $\wedge$ and $\vee$: that M is a sublattice of L means that for all $x, y \in M$, $x \wedge y$ and $x \vee y$, when calculated in L, are elements of M. A sublattice is therefore something more than a subset with the induced order; see the following examples.

Example 7.1.1. The space of real-valued continuous functions on a topological space is a lattice with the usual order: $f \leqslant g$ if and only if $f(x) \leqslant g(x)$ for all x. The space $C^1(\mathbf{R}^n)$ of continuously differentiable functions on $\mathbf{R}^n$ is not a sublattice of $C(\mathbf{R}^n)$. It is not even a lattice on its own.

Example 7.1.2. The family $\mathscr{P}(W)$ of all subsets of a set W is a complete lattice, with $\bigwedge A_j = \bigcap A_j$ and $\bigvee A_j = \bigcup A_j$. The compact sets in $\mathbf{R}^n$ form a sublattice $\mathscr{C}(\mathbf{R}^n)$ of $\mathscr{P}(\mathbf{R}^n)$. This lattice is a complete inf-semilattice but not a complete sup-semilattice. The family $\mathscr{K}(\mathbf{R}^n)$ of all convex compact sets is a lattice but not a sublattice of $\mathscr{C}(\mathbf{R}^n)$: the supremum of two compact sets is not the same in the two lattices.

Example 7.1.3. The family of all closed sets in $\mathbf{R}^n$, denoted by $\mathscr{F}(\mathbf{R}^n)$, is a sublattice of $\mathscr{P}(\mathbf{R}^n)$: the union and intersection of two closed sets are closed. But, although $\mathscr{F}(\mathbf{R}^n)$ is a complete lattice, it is not a sub-complete-lattice of the complete lattice $\mathscr{P}(\mathbf{R}^n)$. The union of a family of closed sets is not always closed, but there is a supremum, viz. the closure of the union. Thus, finite suprema agree with those in $\mathscr{P}(\mathbf{R}^n)$ while infinite suprema do not. This example shows that we would need a different word for *complete lattice*, to allow for a better term than *sub-complete-lattice.*

Example 7.1.4. The set $[-\infty, +\infty]^{\mathbf{R}^n}$ of all functions defined on $\mathbf{R}^n$ and with values in the extended real line is a lattice under the usual order for real numbers, extended in an obvious way to the two infinities. The subset of all convex functions is ordered in the same way, and is also a lattice under this order. However, the convex functions $CVX(\mathbf{R}^n)$ do not form a sublattice of $[-\infty, +\infty]^{\mathbf{R}^n}$. The supremum of two convex functions is equal to the pointwise supremum of them:

$$f \vee g = \max(f, g),$$

but the infima are different in the two lattices: the infimum in the lattice of convex functions is

$$f \wedge_{cvx} g = \sup \big[h \in CVX(\mathbf{R}^n); h \leqslant f, g\big] \leqslant \min(f, g),$$

where the supremum is calculated in $[-\infty, +\infty]^{\mathbf{R}^n}$ and has a sense because that lattice is complete. That the two infima may be different is shown by easy examples like $f(x) = e^x$, $g(x) = e^{-x}$, $x \in \mathbf{R}$. Here $\min(f, g)(x) = e^{-|x|}$, $f \wedge_{cvx} g = 0$.

7.2. Morphology on lattices

We already defined the epigraph of a mapping $X \to [-\infty, +\infty]$; see (3.2.7). The definition makes of course sense for any mapping $X \to Y$, where X is an arbitrary set and Y a preordered set, thus

(7.2.1) $$\operatorname{epi} f = \{(x, y) \in X \times Y; f(x) \leqslant y\}.$$

It will be convenient to define also the *hypograph* of a mapping $f\colon X \to Y$; it is

(7.2.2) $$\operatorname{hypo} f = \{(x, y) \in X \times Y; y \leqslant f(x)\}.$$

The intersection of the two, $\operatorname{epi} f \cap \operatorname{hypo} f$, contains the *graph* of f, which is the set of all pairs $(x, f(x))$:

(7.2.3) $$\operatorname{graph} f = \{(x, y) \in X \times Y; y = f(x)\}.$$

If Y is not only preordered but ordered, the intersection is equal to the graph. If Y has the discrete order, then $\operatorname{epi} f = \operatorname{hypo} f = \operatorname{graph} f$; if Y has the chaotic preorder, then $\operatorname{epi} f = \operatorname{hypo} f = X \times Y$.

In $\mathsf{Incr}(L, L)$, the set of all increasing mappings of an ordered set into itself, the idempotent mappings are of particular interest. We shall call an increasing and idempotent mapping from L into L a *morphological filter*, and denote the set of all such mappings by $\mathsf{Filt}(L)$. Openings and closings are examples: an opening is an antiextensive mapping in $\mathsf{Filt}(L)$, and a closing is an extensive mapping in $\mathsf{Filt}(L)$.

We have studied dilations in an abelian group, i.e., mappings of the form $A \mapsto A + B$. We also remarked (Proposition 3.1.4) that a mapping which commutes with translations and the formation of infinite unions is necessarily of this form. In lattice theory it is therefore natural to say that a mapping $\delta\colon L \to M$, where L and M are complete lattices, is a *dilation* if it commutes with the formation of suprema, i.e., $\delta\big(\bigvee_j x_j\big) = \bigvee_j \delta(x_j)$ for all indexed families $(x_j)_{j\in J}$. In particular we get $\delta(\mathbf{0}_L) = \mathbf{0}_M$ (take J empty), while $\delta(\mathbf{1}_L) = \bigvee_{x\in L} \delta(x) \leqslant \mathbf{1}_M$ (give an example where $\delta(\mathbf{1}_L) \neq \mathbf{1}_M$). Similarly we shall say that ε is an *erosion* if it commutes with the formation of infinite intersections, $\varepsilon\big(\bigwedge_j x_j\big) = \bigwedge_j \varepsilon(x_j)$. We note that $\varepsilon(\mathbf{1}_L) = \mathbf{1}_M$ (take J empty), while $\varepsilon(\mathbf{0}_L) = \bigwedge_{x\in L} \varepsilon(x) \geqslant \mathbf{0}_M$.

Dilations and erosions are always increasing. Indeed, we have $\delta(x \vee y) = \delta(x) \vee \delta(y)$. If $x \leqslant_L y$, this equation simplifies to $\delta(y) = \delta(x) \vee \delta(y) \geqslant_M \delta(x)$, which shows that δ is increasing, $\delta \in \mathsf{Incr}(L, M)$. A similar argument shows that erosions are increasing.

We now recall Proposition 3.1.3, where it was stated that, in an abelian group, $\delta_B(A) \subset C$ if and only if $A \subset \varepsilon_B(C)$. In a lattice this may be written as $\delta(x) \leqslant y$ iff $x \leqslant \varepsilon(y)$, equivalently as $\operatorname{epi}\delta = (\operatorname{hypo}\varepsilon)^{-1}$, where the exponent -1 means that we swap the components: for a subset A of a Cartesian product $X \times Y$ we define $A^{-1} = \{(y,x); (x,y) \in A\} \subset Y \times X$. May we use this as a model to define erosions from dilations and conversely in the more general lattice situation? Indeed this is the case, and we shall now show this.

7.3. Inverses of mappings between lattices

Let L be a complete lattice, M a preordered set, and $f\colon L \to M$ any mapping. We then define the *upper inverse* $f^{[-1]}\colon M \to L$ and the *lower inverse* $f_{[-1]}\colon M \to L$ as the mappings

$$f^{[-1]}(y) = \bigwedge_{x \in L}(x; f(x) \geqslant_M y) = \bigwedge_{x \in L}(x; (x,y) \in \operatorname{hypo} f), \qquad y \in M; \tag{7.3.1}$$

$$f_{[-1]}(y) = \bigvee_{x \in L}(x; f(x) \leqslant_M y) = \bigvee_{x \in L}(x; (x,y) \in \operatorname{epi} f), \qquad y \in M. \tag{7.3.2}$$

As a first observation, let us note that these inverses are always increasing. If there exists a smallest element $\mathbf{0}_M$, then $f^{[-1]}(\mathbf{0}_M) = \mathbf{0}_L$. Similarly, if M possesses a largest element $\mathbf{1}_M$, then $f_{[-1]}(\mathbf{1}_M) = \mathbf{1}_L$. If M has the chaotic preorder, then both inverses are constant, $f^{[-1]} = \mathbf{0}_L$ and $f_{[-1]} = \mathbf{1}_L$ identically.

We note that we always have

$$\operatorname{hypo} f \subset \left(\operatorname{epi} f^{[-1]}\right)^{-1} \text{ and } \operatorname{epi} f \subset \left(\operatorname{hypo} f_{[-1]}\right)^{-1}. \tag{7.3.3}$$

If, given a mapping $f\colon L \to M$, we can find a mapping $g\colon M \to L$ such that $\operatorname{epi} g = (\operatorname{hypo} f)^{-1}$ we would be content to have a kind of inverse to f. However, usually the best we can do is to study mappings with $\operatorname{epi} g \supset (\operatorname{hypo} f)^{-1}$ or $\operatorname{epi} h \subset (\operatorname{hypo} f)^{-1}$. This we shall do in the following proposition, which shows that the upper and lower inverses are solutions to certain extremal problems.

Proposition 7.3.1. *Let L be a complete lattice, M a preordered set, and let $f\colon L \to M$, $g,h\colon M \to L$ be mappings. If* $\operatorname{epi} g \supset (\operatorname{hypo} f)^{-1} \supset \operatorname{epi} h$, *then $g \leqslant f^{[-1]} \leqslant h$ and*

$$\operatorname{epi} g \supset \operatorname{epi} f^{[-1]} \supset (\operatorname{hypo} f)^{-1} \supset \operatorname{epi} h.$$

Hence $f^{[-1]}$ is the largest mapping g such that $\operatorname{epi} g \supset (\operatorname{hypo} f)^{-1}$. *If on the other hand* $\operatorname{hypo} g \subset (\operatorname{epi} f)^{-1} \subset \operatorname{hypo} h$, *then $g \leqslant f_{[-1]} \leqslant h$ and*

$$\operatorname{hypo} g \subset (\operatorname{epi} f)^{-1} \subset \operatorname{hypo} f_{[-1]} \subset \operatorname{hypo} h.$$

Hence $f_{[-1]}$ is the smallest mapping h which satisfies $\operatorname{hypo} h \supset (\operatorname{epi} f)^{-1}$.

The proof is straightforward.

Corollary 7.3.2. *With f, g and h given as in the proposition, assume that* $\operatorname{epi} g = (\operatorname{hypo} f)^{-1}$. *Then $g = f^{[-1]}$. Similarly, if* $\operatorname{hypo} h = (\operatorname{epi} f)^{-1}$, *then $h = f_{[-1]}$. If also M is a complete lattice, then* $\operatorname{epi} g = (\operatorname{hypo} f)^{-1}$ *implies that $g_{[-1]} = f$ in addition to $g = f^{[-1]}$. Similarly,* $\operatorname{hypo} h = (\operatorname{epi} f)^{-1}$ *implies $h^{[-1]} = f$ in addition to $h = f_{[-1]}$.*

An ideal inverse g would satisfy $g \circ f = \mathsf{Id}_L$, $f \circ g = \mathsf{Id}_M$, and the inverse of g would be f. It is therefore natural to compare $f^{[-1]} \circ f$ and $f_{[-1]} \circ f$ with Id_L; $f \circ f^{[-1]}$ and $f \circ f_{[-1]}$ with Id_M; and $(f_{[-1]})^{[-1]}$ and $(f^{[-1]})_{[-1]}$ with f. This is what we shall do now.

Proposition 7.3.3. *If L is a complete lattice and M a preordered set, then for all mappings $f\colon L \to M$ one has $f^{[-1]} \circ f \leqslant \mathsf{Id}_L \leqslant f_{[-1]} \circ f$ with equality if f is coincreasing in the sense of Definition 2.1.5. Conversely, if $f^{[-1]} \circ f = \mathsf{Id}_L$ or $f_{[-1]} \circ f = \mathsf{Id}_L$, then f is coincreasing.*

Proof. It is clear that $f^{[-1]}(f(a)) = \bigwedge(x; f(x) \geqslant f(a)) \leqslant a$. If f is coincreasing, then $\{x; f(x) \geqslant f(a)\}$ is contained in $\{x; x \geqslant a\}$, so that $f^{[-1]}(f(a)) \geqslant \bigwedge(x; x \geqslant a) = a$.

Conversely, if $f^{[-1]}(f(a)) \geqslant a$, then for all x, $f(x) \geqslant f(a)$ implies $x \geqslant a$. If this is true for all a, then f is coincreasing.

Corollary 7.3.4. *Let L be a complete lattice and M a preordered set. Then $f^{[-1]}(y) \leqslant f_{[-1]}(y)$ for all $y \in \operatorname{im} f$, and also for all y majorizing or minorizing $\operatorname{im} f$. In particular, $f^{[-1]} \leqslant f_{[-1]}$ if f is surjective.*

Proof. The statement for $y \in \operatorname{im} f$ follows directly from the proposition. If y majorizes all elements in $\operatorname{im} f$, then $f_{[-1]}(y) = \mathbf{1}$, and if y minorizes all elements in $\operatorname{im} f$, then $f^{[-1]}(y) = \mathbf{0}$.

Proposition 7.3.5. *If g, h are increasing mappings such that $g \circ f \leqslant \mathsf{Id}_L \leqslant h \circ f$, then $g \leqslant f^{[-1]}$ and $h \geqslant f_{[-1]}$. Hence, in view of Proposition 7.3.3, $f^{[-1]}$ is the largest increasing mapping g such that $g \circ f \leqslant \mathsf{Id}_L$, and $f_{[-1]}$ is the smallest increasing mapping h such that $h \circ f \geqslant \mathsf{Id}_L$.*

Proof. If g and h are increasing and $g \circ f \leqslant \mathsf{Id}_L \leqslant h \circ f$, then $\operatorname{epi} g \supset (\operatorname{hypo} f)^{-1}$ and $\operatorname{hypo} h \supset (\operatorname{epi} f)^{-1}$. We can now apply Proposition 7.3.1.

Proposition 7.3.6. *Let L be a complete lattice and M a preordered set. Then the following four conditions are equivalent.*

(a) *f is coincreasing;*

(b) *$f^{[-1]} \circ f \geqslant \mathsf{Id}_L$;*

(c) *$f_{[-1]} \circ f \leqslant \mathsf{Id}_L$;*

(d) *$f_{[-1]} \leqslant f^{[-1]}$.*

Proof. (a) implies (b) and (c). If f is coincreasing we know already from Proposition 7.3.3 that (b) and (c) hold with equality.

(b) or (c) implies (a). If (b) or (c) holds, then, in view of Proposition 7.3.3, they hold with equality, and f is coincreasing.

(a) implies (d). Assume that f is coincreasing and fix an element $y \in M$. Let $x, x' \in L$ be such that $f(x) \leqslant y \leqslant f(x')$. Then $x \leqslant x'$. Letting x vary we see that $f_{[-1]}(y) \leqslant x'$. Letting now also x' vary, we see that $f_{[-1]}(y) \leqslant f^{[-1]}(y)$. Thus (d) holds.

(d) implies (a). If x and x' are given with $f(x) \leqslant f(x')$ we define $y = f(x)$. Then $x \leqslant f_{[-1]}(y)$ and $f^{[-1]}(y) \leqslant x'$. If we know that $f_{[-1]}(y) \leqslant f^{[-1]}(y)$ it follows that $x \leqslant x'$, proving that f is coincreasing.

Next we compose $f_{[-1]}$ with f in the other order. This will lead to a characterization of dilations—and, by duality, of erosions.

Theorem 7.3.7. *If L and M are complete lattices and $f\colon L \to M$ any mapping, then the following five properties are equivalent.*
(A) *f is a dilation;*
(B) $\operatorname{epi} f \supset (\operatorname{hypo} f_{[-1]})^{-1}$*;*
(C) $\operatorname{epi} f = (\operatorname{hypo} f_{[-1]})^{-1}$*;*
(D) *f is increasing and* $\operatorname{epi} f \supset (\operatorname{graph} f_{[-1]})^{-1}$*;*
(E) *f is increasing and $f \circ f_{[-1]} \leqslant \mathsf{Id}_M$.*

Proof. (A) implies (B). Suppose that (A) holds. Then if $(y,x) \in \operatorname{hypo} f_{[-1]}$, in other words if $x \leqslant f_{[-1]}(y)$, we obtain, since f is increasing by hypothesis,

$$f(x) \leqslant f(f_{[-1]}(y)) = f\big(\bigvee(x; f(x) \leqslant y)\big) = \bigvee \big(f(x); f(x) \leqslant y\big) \leqslant y,$$

which means that $(x,y) \in \operatorname{epi} f$. Thus (B) holds.

(B) implies (A). We note first that f is increasing if (B) holds. Indeed, if $x \leqslant x'$ and we define $y = f(x')$, then $f_{[-1]}(y) = f_{[-1]}(f(x')) \geqslant x' \geqslant x$ (see Proposition 7.3.3), which by (B) implies that $f(x) \leqslant y = f(x')$.

Let now any family $(x_j)_j$ of elements of L be given and define $z = \bigvee f(x_j)$, $w = f(\bigvee x_j)$. Since f is increasing we always have $z \leqslant w$. Is it true that $w \leqslant z$? We note that, by Proposition 7.3.3, $x_j \leqslant f_{[-1]}(f(x_j)) \leqslant f_{[-1]}(z)$. Taking the supremum over all j we obtain $\bigvee x_j \leqslant f_{[-1]}(z)$, which by (B) implies that $w = f(\bigvee x_j) \leqslant z$. We have proved (A).

(B) is equivalent to (C). This is clear since we always have $\operatorname{epi} f \subset (\operatorname{hypo} f_{[-1]})^{-1}$.

(B) implies (D). We have seen that (B) implies that f is increasing. That (B) implies $(\operatorname{graph} f_{[-1]})^{-1} \subset \operatorname{epi} f$ follows from the inclusion $\operatorname{graph} f_{[-1]} \subset \operatorname{hypo} f_{[-1]}$.

(D) implies (B). If $x \leqslant f_{[-1]}(y)$ we define $x' = f_{[-1]}(y)$ and note that $x \leqslant x'$ and that $(y, x') \in \operatorname{graph} f_{[-1]}$. If (D) holds, we conclude that $f(x') \leqslant y$. Hence, if f is increasing, $f(x) \leqslant f(x') \leqslant y$, proving (B).

(D) and (E) are equivalent: (E) is just a rephrasing of (D).

Remark. We may use (B) or (C) to define dilations $L \to M$ when M is only a preordered set.

Corollary 7.3.8. *If L and M are complete lattices and $f\colon L \to M$ and $g\colon L \to M$ are two mappings such that* $\operatorname{epi} f = (\operatorname{hypo} g)^{-1}$*, then f is a dilation and g is an erosion, and $f_{[-1]} = g$, $g^{[-1]} = f$.*

Proof. It follows from $\operatorname{epi} f = (\operatorname{hypo} g)^{-1}$ that f is increasing and that $f_{[-1]} = g$, hence that (D) in the theorem holds. Since (D) is equivalent to (A), we see that f is a dilation. The rest follows by duality.

Proposition 7.3.9. *If L and M are complete lattices and $f\colon L \to M$ any mapping, then quite generally $\big(f_{[-1]}\big)^{[-1]} \leqslant f \leqslant \big(f^{[-1]}\big)_{[-1]}$. Equality holds at the first place if and only if f is a dilation; at the second place if and only if f is an erosion.*

Proof. We always have $\operatorname{epi} f \subset (\operatorname{hypo} f_{[-1]})^{-1}$, i.e., $y \geqslant f(a)$ implies $f_{[-1]}(y) \geqslant a$. This yields

$$(f_{[-1]})^{[-1]}(a) = \bigwedge(y; f_{[-1]}(y) \geqslant a) \leqslant \bigwedge(y; y \geqslant f(a)) = f(a).$$

If f is a dilation, then, as the last theorem shows, $\operatorname{epi} f = \big(\operatorname{hypo} f_{[-1]}\big)^{-1}$ and equality follows.

Conversely, let us note that, in view of (7.3.3) we always have

$$\operatorname{epi} f \subset \big(\operatorname{hypo} f_{[-1]}\big)^{-1} \subset \operatorname{epi} \big(f_{[-1]}\big)^{[-1]}.$$

Now if $\big(f_{[-1]}\big)^{[-1]} = f$, then these inclusions are equalities, and we conclude that $\operatorname{epi} f = (\operatorname{hypo} f_{[-1]})^{-1}$, which according to Theorem 7.3.7 means that f is a dilation.

Theorem 7.3.10. *If L and M are complete lattices and $\delta\colon L \to M$ a dilation, then $\delta_{[-1]}\colon M \to L$ is an erosion. Similarly, if $\varepsilon\colon L \to M$ is an erosion, then $\varepsilon^{[-1]}$ is a dilation.*

Proof. We know that $g = \delta_{[-1]}$ is increasing, so we have $g(\bigwedge y_j) \leqslant g(y_k)$ for all k; hence $g(\bigwedge y_j) \leqslant \bigwedge g(y_k)$. We need to prove the opposite inequality, $\bigwedge g(y_k) \leqslant g(y)$, where $y = \bigwedge y_j$. From (E) in Theorem 7.3.7 we learn that $\delta(g(y_k)) \leqslant y_k$ for all k, which implies that $\delta(\bigwedge g(y_j)) \leqslant y_k$, hence that $\delta(\bigwedge g(y_j)) \leqslant y$. Since g is increasing we get

$$\bigwedge g(y_j) \leqslant g\left(\delta\left(\bigwedge g(y_j)\right)\right) \leqslant g(y),$$

where the first inequality follows from Proposition 7.3.3. This completes the proof.

Corollary 7.3.11. *For any dilation $\delta\colon L \to M$ we have $\delta \circ \delta_{[-1]} \circ \delta = \delta$ and $\delta_{[-1]} \circ \delta \circ \delta_{[-1]} = \delta_{[-1]}$. Dually $\varepsilon \circ \varepsilon^{[-1]} \circ \varepsilon = \varepsilon$ and $\varepsilon^{[-1]} \circ \varepsilon \circ \varepsilon^{[-1]} = \varepsilon^{[-1]}$ for any erosion $\varepsilon\colon L \to M$. In particular, $\delta_{[-1]} \circ \delta$ and $\delta \circ \delta_{[-1]}$ are idempotent: $\delta_{[-1]} \circ \delta \in \mathsf{Filt}(L)$ and $\delta \circ \delta_{[-1]} \in \mathsf{Filt}(M)$. The first is a closing in L, the second an opening in M. Also $\varepsilon^{[-1]} \circ \varepsilon$ and $\varepsilon \circ \varepsilon^{[-1]}$ are idempotent; the first an opening, the second a closing.*

Proof. We always have $f_{[-1]} \circ f \geqslant \mathsf{Id}_L$ (Proposition 7.3.3); it follows that $f_{[-1]} \circ f \circ f_{[-1]} \geqslant f_{[-1]}$. If f is increasing, we also get $f \circ f_{[-1]} \circ f \geqslant f$.

For dilations we have $\delta \circ \delta_{[-1]} \leqslant \mathsf{Id}_M$ (Theorem 7.3.7), from which we deduce that $\delta \circ \delta_{[-1]} \circ \delta \leqslant \delta$ and $\delta_{[-1]} \circ \delta \circ \delta_{[-1]} \leqslant \delta_{[-1]}$. This shows what we want for dilations; the rest follows by duality.

7.4. Division of mappings between lattices

We shall now generalize the definitions of upper and lower inverses. Let three sets L, M, N be given, and let us assume that M is a complete lattice and N a preordered set. (We shall often assume that all three are complete lattices, but this is not necessary for the definitions to make sense.) Let also two mappings $f\colon L \to M$ and $g\colon L \to N$ be given. Then we may define two mappings $f/^{\star}g, f/_{\star}g\colon N \to M$ by

$$(f/^{\star}g)(y) = \bigwedge_{x \in L}(f(x); g(x) \geqslant_N y), \qquad y \in N, \tag{7.4.1}$$

$$(7.4.2) \qquad (f/_{\star}g)(y) = \bigvee_{x\in L}(f(x); g(x) \leqslant_N y), \qquad y \in N.$$

We shall call them the *upper* and *lower quotient* of f and g.

$$\begin{tikzcd}
N \arrow[r, "f/^{\star}g{,}f/_{\star}g"] & M \\
L \arrow[u, "g"] \arrow[r, equal] & L \arrow[u, "f"']
\end{tikzcd}$$

The quotients $f/^{\star}g$ and $f/_{\star}g$ increase when f increases and they decrease when g increases—just as with division of positive numbers,

(7.4.3) $f_1 \leqslant_M f_2$ and $g_1 \geqslant_N g_2$ implies that $f_1/^{\star}g_1 \leqslant_M f_2/^{\star}g_2$ and $f_1/_{\star}g_1 \leqslant_M f_2/_{\star}g_2$.

The mappings $f/^{\star}g$ and $f/_{\star}g$ are always increasing. If $g(x) \geqslant_N y$, then $f(x) \geqslant_M (f/^{\star}g)(y)$; if $g(x) \leqslant_N y$, then $f(x) \leqslant_M (f/_{\star}g)(y)$. In particular, if $g(x) = y$, then $(f/^{\star}g)(y) \leqslant_M f(x) \leqslant_M (f/_{\star}g)(y)$.

If we specialize the definitions to the situation when $L = M$ and $f = \mathsf{Id}_L$, then $f/^{\star}g = \mathsf{Id}_L/^{\star}g = g^{[-1]}$ and $f/_{\star}g = \mathsf{Id}_L/_{\star}g = g_{[-1]}$; cf. (7.3.1) and (7.3.2).

On the other hand, if we specialize to the situation where L is an arbitrary subset of a complete lattice M, $N = M$, and g is the inclusion mapping $L \to M$, then $f/^{\star}g = f^{\diamond}$ and $f/_{\star}g = f_{\diamond}$, where we define $f^{\diamond}$ and $f_{\diamond}$ as in (4.3.2), generalized to any complete lattice. If we specialize further, letting also f be the inclusion mapping, we obtain

$$(f/_{\star}g)(y) = (f/_{\star}f)(y) = f_{\diamond}(y) = \bigvee_{x\in L}(x; x \leqslant y) = y^{\circ} \in M.$$

It is easy to verify that $y \mapsto y^{\circ}$ is an opening. The elements y such that $y^{\circ} = y$ are called *L-convex*. The reason should be clear from the following example.

Example 7.4.1. Le M be the complete lattice $[-\infty, +\infty]^E$ of functions on a vector space E, let F be a vector subspace of its dual $E^{\star}$, and let L be the set of affine functions with linear part in F, i.e., functions of the form $\alpha(x) = \xi(x) + c$ for some linear form $\xi \in F$ and some real constant c. Then a function on E is L-convex if and only if it is equal to the supremum of all its affine minorants belonging to L. By Fenchel's theorem, Theorem 5.6.9, this happens if and only if it is convex in the usual sense, lower semicontinuous for the topology $\sigma(E, F)$ on E generated by the linear forms in F, and not taking the value $-\infty$ except when it is equal to the constant $-\infty$.

Proposition 7.4.2. *If $f\colon L \to M$ is increasing and $g\colon L \to N$ is coincreasing, then $f/_{\star}g \leqslant f/^{\star}g$. If f is arbitrary and $y \in \operatorname{im} g$, then $(f/^{\star}g)(y) \leqslant (f/_{\star}g)(y)$. In particular, if g is surjective and f is arbitrary, then $f/^{\star}g \leqslant f/_{\star}g$.*

The upper quotient $f/^{\star}g$ is the optimal solution to an inequality:

Proposition 7.4.3. *For all mappings $f\colon L \to M$ and $g\colon L \to N$ we have*

$$(7.4.4) \qquad (f/^{\star}g) \circ g \leqslant f \leqslant (f/_{\star}g) \circ g,$$

with equality if f is increasing and g is coincreasing. Conversely, if $h, k \in \mathsf{Incr}(N, M)$ are two increasing functions such that $h \circ g \leqslant f \leqslant k \circ g$, then $h \leqslant f/^{\star}g$ and $k \geqslant f/_{\star}g$. Thus $f/^{\star}g$ is the largest increasing function h such that $h \circ g \leqslant f$, and $f/_{\star}g$ is the smallest increasing function k such that $f \leqslant k \circ g$. In the special case $L = N$ and $g = \mathsf{Id}_L$ we obtain

$$f/^{\star}\mathsf{Id}_L \leqslant f \leqslant f/_{\star}\mathsf{Id}_L,$$

where $f/^{\star}\mathsf{Id}_L$ is the largest increasing minorant of f and $f_{\star}/\mathsf{Id}_L$ is the smallest increasing majorant of f; when f itself is increasing we therefore get equality.

We next compare the quotient $f/^{\star}g$ and the composition $f \circ g^{[-1]}$:

Proposition 7.4.4. *For all increasing mappings $f\colon L \to M$ and all mappings $g\colon L \to N$ we have $f/^{\star}g \geqslant f \circ g^{[-1]}$ with equality if f is an erosion, and $f/_{\star}g \leqslant f \circ g_{[-1]}$ with equality if f is a dilation. If g is coincreasing, then $f/_{\star}g \leqslant f \circ g_{[-1]} \leqslant f \circ g^{[-1]} \leqslant f/^{\star}g$.*

Proposition 7.4.5. *If $h \in \mathsf{Incr}(M, P)$, where P is an ordered set, we have $h \circ (f/^{\star}g) \leqslant (h \circ f)/^{\star}g$ with equality if h is an erosion. Similarly $h \circ (f/_{\star}g) \geqslant (h \circ f)/_{\star}g$ with equality if h is a dilation. A special case is $h \circ (f/^{\star}\mathsf{Id}_L) \leqslant (h \circ f)/^{\star}\mathsf{Id}_L$ (take $L = N$ and $g = \mathsf{Id}_L$). Another special case is Proposition 7.4.4 (take $L = M$ and $f = \mathsf{Id}_L$).*

Proposition 7.4.6. *For all mappings $f\colon L \to M$ we have*

$$f/_{\star}f \leqslant \mathsf{Id}_M \leqslant f/^{\star}f. \tag{7.4.5}$$

Corollary 7.4.7. *For all mappings $f\colon L \to M$ we have*

$$(f/_{\star}f) \circ f = f = (f/^{\star}f) \circ f. \tag{7.4.6}$$

Proof. The result follows on combining (7.4.4), taking $g = f$, and (7.4.5), multiplied from the right by f—or directly from the definitions.

Theorem 7.4.8. *Let $f\colon L \to M$ be any mapping from a set L into a complete lattice M. Then $\eta = f/_{\star}f\colon M \to M$ is an opening. Conversely, any opening in M is of this form for some mapping $f\colon L \to M$ with $L = M$.*

Proof. It is clear that $\eta(y) = \bigvee(f(x); f(x) \leqslant y)$ defines a mapping $M \to M$ which is increasing and antiextensive. The latter property implies that $\eta \circ \eta \leqslant \eta$. It remains to be proved that $\eta \leqslant \eta \circ \eta$. To do so we note that, by definition, $f(x) \leqslant y$ implies $f(x) \leqslant \eta(y)$. Therefore

$$\eta(y) = \bigvee(f(x); f(x) \leqslant y) \leqslant \bigvee(f(x); f(x) \leqslant \eta(y)) = \eta(\eta(y)),$$

proving that $\eta \leqslant \eta \circ \eta$. Note that, by (7.4.6), $\eta \circ f = f$, proving that the image of f is a subset of Inv_η; in other words, all elements $f(x)$ are η-open.

The last statement follows from the formula $\eta /_{\!\star}\, \eta = \eta$, which holds for any opening $\eta\colon M \to M$ and is straightforward to prove. Writing out the formula in full if $\eta = f /_{\!\star}\, f$, we obtain $(f /_{\!\star}\, f) /_{\!\star}\, (f /_{\!\star}\, f) = f /_{\!\star}\, f$.

7.5. Structure theorems in lattice morphology

Lemma 7.5.1. *Let L and M be two complete lattices and let $a \in L$ and $b, c \in M$ be three fixed elements. Define a mapping $\varepsilon = \varepsilon_a^{b,c}\colon L \to M$ by*

$$\varepsilon_a^{b,c}(x) = \begin{cases} \mathbf{1}_M, & x = \mathbf{1}_L, \\ c, & x \neq \mathbf{1}_L,\ x \geqslant a, \\ b, & x \ngeqslant a. \end{cases}$$

If $b \leqslant c$, then $\varepsilon_a^{b,c}$ is an erosion.

Proof. We have to prove that $\varepsilon\big(\bigwedge_{j\in J} x_j\big)$ and $\bigwedge_{j\in J} \varepsilon(x_j)$ are equal. If $x_j = \mathbf{1}_L$ for all $j \in J$ (thus in particular if J is empty), then both expressions take the value $\mathbf{1}_M$; if for all j we have $x_j \geqslant a$ and there is a j such that $x_j \neq \mathbf{1}_L$, then both expressions take the value c; finally, if there is j such that $x_j \ngeqslant a$, then $\varepsilon(\bigwedge x_j) = b$ and $\varepsilon(x_j)$ may be either b or c or $\mathbf{1}_M$, but $\varepsilon(x_j) = b$ for at least one j, so if $b \leqslant c$ we have $\bigwedge \varepsilon(x_j) = b$ and equality holds.

We can now prove a structure theorem for increasing mappings which is analogous to Matheron's first structure theorem, Theorem 4.3.1; cf. Serra (1988:20, 2001:29), where the result is given for $L = M$.

Theorem 7.5.2. *Let $f\colon L \to M$ be an increasing mapping from a complete lattice L into another, M. Then f is the supremum of a family of elementary erosions as defined in Lemma 7.5.1; more precisely,*

$$f = \bigvee_{a\in L} \varepsilon_a^{0,f(a)}.$$

Proof. On the one hand we have $\varepsilon_a^{0,f(a)}(x) \leqslant f(x)$ for all $a, x \in L$. On the other hand $\varepsilon_a^{0,f(a)}(a) = f(a)$ for every $a \in L$. Hence the equality.

Definition 7.5.3. *Given two complete lattices L and M and two elements $a \in L$, $b \in M$ we define mappings $\varepsilon_a^b\colon L \to M$, $\delta_b^a\colon M \to L$, $\eta_a\colon L \to L$, and $\gamma_b\colon M \to M$ by*

$$\varepsilon_a^b(x) = \begin{cases} \mathbf{1}_M, & x \geqslant a, \\ b, & x \ngeqslant a; \end{cases} \qquad \delta_b^a(y) = \begin{cases} \mathbf{0}_L, & y \leqslant b, \\ a, & y \nleqslant b; \end{cases}$$

$$\eta_a(x) = \begin{cases} a, & x \geqslant a, \\ \mathbf{0}_L, & x \ngeqslant a; \end{cases} \qquad \gamma_b(y) = \begin{cases} b, & y \leqslant b, \\ \mathbf{1}_M, & y \nleqslant b. \end{cases}$$

Note that $\varepsilon_a^b = \varepsilon_a^{b,1}$ in the notation of Lemma 7.5.1.

Lemma 7.5.4. $(\varepsilon_a^b)^{[-1]} = \delta_b^a$.

Proof. Let us recall that the upper inverse is defined by $(\varepsilon_a^b)^{[-1]}(y) = \bigwedge_x \bigl(x; \varepsilon_a^b(x) \geqslant y\bigr)$. If $b \geqslant y$, then $\varepsilon_a^b(x) \geqslant y$ for all $x \in L$, so the upper inverse of ε_a^b takes the value $\mathbf{0}_L$ at y, just as does $\delta_b^a(y)$. If on the other hand $b \ngeqslant y$, then $\varepsilon_a^b(x) \geqslant y$ if and only if $x \geqslant a$, so the upper inverse of ε_a^b takes the value a at y; so does δ_b^a.

Lemma 7.5.5. *For $a \neq \mathbf{0}$ and all b we have $\varepsilon_a^b \circ \delta_b^a = \gamma_b$. For $a = \mathbf{0}$ and all b we have $\varepsilon_a^b \circ \delta_b^a = \varepsilon_{\mathbf{0}}^b \circ \delta_b^{\mathbf{0}} = \gamma_{\mathbf{1}}$. For all a and all $b \neq \mathbf{1}$ we have $\delta_b^a \circ \varepsilon_a^b = \eta_a$. For all a and $b = \mathbf{1}$ we have $\delta_b^a \circ \varepsilon_a^b = \delta_{\mathbf{1}}^a \circ \varepsilon_a^{\mathbf{1}} = \eta_{\mathbf{0}}$.*

Proof. If $a \neq \mathbf{0}$, then $\varepsilon(\delta(y)) = \varepsilon(\mathbf{0}) = b$ if $y \leqslant b$, and $\varepsilon(\delta(y)) = \varepsilon(a) = \mathbf{1}$ otherwise. These are the same values as $\gamma_b(y)$. If on the other hand $a = \mathbf{0}$ we have $\varepsilon(\delta(y)) = \varepsilon(\mathbf{0}) = \mathbf{1} = \gamma_{\mathbf{1}}(y)$ for all b. The proof for the other composition follows by duality.

We can now prove a structure theorem for openings which is analogous to Matheron's second structure theorem, Theorem 4.3.5; cf. Serra (1988:22, 2001:49).

Theorem 7.5.6. *If $\theta\colon L \to L$ is an opening in a complete lattice L, then it is the supremum of a family of elementary openings η_a as defined in Definition 7.5.3:*

$$\theta = \bigvee_{a \in A} \eta_a,$$

where $A = \mathsf{Inv}_\theta$ is the invariance set of θ, the set of all θ-open elements. Conversely, if A is an arbitrary subset of L, then this formula defines an opening. Its invariance set is the set generated by A under the formation of suprema.

Proof. We know from Proposition 4.3.4 that

$$\theta(x) = (\mathsf{Id}_{\mathsf{Inv}_\theta})_\diamond(x) = \bigvee_{a \in A} \bigl(a; a \leqslant x\bigr).$$

We claim that the last element is equal to $\bigvee_{a \in A} \eta_a(x)$. Indeed, since $a = \eta_a(a) = \eta_a(x)$ for all $a \leqslant x$, we get

$$\bigvee_{a \in A} \bigl(a; a \leqslant x\bigr) = \bigvee_{a \in A} \bigl(\eta_a(a); a \leqslant x\bigr) = \bigvee_{a \in A} \bigl(\eta_a(x); a \leqslant x\bigr) = \bigvee_{a \in A} \eta_a(x),$$

where the last equality follows from the fact that $\eta_a(x) = \mathbf{0}$ when $a \nleqslant x$ so that these a do not contribute to the supremum.

7.6. Strong filters in lattices

If L is a lattice we shall denote by $L' = \mathsf{Incr}(L, L)$ the set of all increasing mappings of L into itself. We may then form $L'' = \mathsf{Incr}(L', L')$; the mapping $f \mapsto f \circ f$ is an example of an element of L''.

The order in L' being the obvious one, we note that L' itself is a complete lattice if L is a complete lattice, and that we have

$$(\bigvee g_j) \circ f = \bigvee (g_j \circ f) \text{ as well as } (\bigwedge g_j) \circ f = \bigwedge (g_j \circ f). \tag{7.6.1}$$

In particular, taking $g_1 = \mathsf{Id}$ and $g_2 = f$, we note that

$$(\mathsf{Id} \vee f) \circ f = f \vee (f \circ f) \text{ and } (\mathsf{Id} \wedge f) \circ f = f \wedge (f \circ f). \tag{7.6.2}$$

We shall write

$$f^{(1)} = \mathsf{Id} \vee f \text{ and } f_{(1)} = \mathsf{Id} \wedge f,$$

so that (7.6.2) can be written

$$f^{(1)} \circ f = f \vee (f \circ f) \text{ and } f_{(1)} \circ f = f \wedge (f \circ f),$$

or more briefly

$$f^{(1)} f = f \vee ff \text{ and } f_{(1)} f = f \wedge ff,$$

if we write composition as juxtaposition.

However, in contrast to (7.6.1), we only have

$$g \circ (\textstyle\bigvee f_j) \geqslant \bigvee (g \circ f_j) \text{ and } g \circ (\bigwedge f_j) \leqslant \bigwedge (g \circ f_j), \tag{7.6.3}$$

where the inequalities can be strict. In particular we have

$$ff^{(1)} \geqslant f \vee ff = f^{(1)} f \text{ and } ff_{(1)} \leqslant f \wedge ff = f_{(1)} f. \tag{7.6.4}$$

Among the increasing mappings, the idempotent ones play an important role. We already introduced the notation $\mathsf{Filt}(L)$ for the set of increasing and idempotent mappings of L into itself. Its elements are called *morphological filters*, or just *filters* when any mixup with the set-theoretical filters seems improbable.

For any filter f we thus have

$$ff_{(1)} \leqslant f_{(1)} f = f = f^{(1)} f \leqslant f \circ f^{(1)}, \qquad f \in \mathsf{Filt}(L). \tag{7.6.5}$$

Definition 7.6.1. *Let $f \in \mathsf{Filt}(L)$. Then f is called a* sup-filter *if $f \circ f^{(1)} = f$, and it is called an* inf-filter *if $f \circ f_{(1)} = f$. If f is both a sup-filter and an inf-filter, we shall say that f is a* strong filter*.*

For a strong filter f we thus have equality in (7.6.5), and we see that a strong filter f commutes with $f^{(1)}$ and $f_{(1)}$. Actually also $f^{(1)}$ and $f_{(1)}$ commute if f is a strong filter (prove this).

An idempotent mapping f is constant on $\{x, f(x)\}$ for every x. A filter is therefore constant on the segment $[x, f(x)]$ if $x \leqslant f(x)$ and on the segment $[f(x), x]$ if $f(x) \leqslant x$. A sup-filter is constant on the segment $[x, f^{(1)}(x)]$, an inf-filter is constant on $[f_{(1)}(x), x]$, and a strong filter on $[f_{(1)}(x), f^{(1)}(x)]$ for every x.

A filter f is a closing if and only if $f^{(1)} = f$ (and if and only if $f_{(1)} = \mathsf{Id}$); it is an opening if and only if $f_{(1)} = f$ (and if and only if $f^{(1)} = \mathsf{Id}$). Every closing and every opening are strong filters. However, there are filters that are not strong:

Example 7.6.2. Let us define $f \colon L \to L$ by

$$f(x) = \begin{cases} x & \text{if } x \geqslant b \text{ or } x \leqslant b, \\ b & \text{otherwise.} \end{cases}$$

Then we can prove that f is increasing and idempotent, thus a filter. But it is not a sup-filter, neither an inf-filter if there exists an x which is not comparable with b. This example is from Matheron (1988:116).

Theorem 7.6.3. *If $\eta\colon L \to L$ is an opening and γ is a closing, then $f = \eta \circ \gamma$ is a sup-filter, and $g = \gamma \circ \eta$ is an inf-filter. Conversely, every sup-filter is of the form $\eta \circ \gamma$ for some opening η and some closing γ.*

Proof. Clearly f is increasing. To prove that it is idempotent we note that

$$f \circ f = (\eta \circ \gamma) \circ (\eta \circ \gamma) \geqslant (\eta \circ \mathsf{Id}) \circ (\eta \circ \gamma) = \eta \circ \eta \circ \gamma = f,$$

and that

$$f \circ f = (\eta \circ \gamma) \circ (\eta \circ \gamma) \leqslant (\eta \circ \gamma) \circ (\mathsf{Id} \circ \gamma) = \eta \circ \gamma \circ \gamma = f.$$

Finally we see that f is a sup-filter from the general inequality $f \circ f^{(1)} \geqslant f$ and

$$f \circ f^{(1)} = (\eta \circ \gamma) \circ (\mathsf{Id} \vee (\eta \circ \gamma)) \leqslant (\eta \circ \gamma) \circ \gamma = \eta \circ \gamma = f.$$

The second part is more difficult to prove, and we leave it for now.

7.7. Exercises

7.1. Let f be a constant mappig $L \to M$, $f(x) = b$, where L and M are complete lattices and b a fixed element of M. Determine the upper and lower inverses of f.

7.2. Let L be a complete lattice and f, g the mappings $L \to L$ given by $f(x) = a \vee x$, $g(x) = a \wedge x$, where a is a fixed element of L.

(a) Determine $f_{[-1]}$ and $g^{[-1]}$.

(b) Determine $f^{[-1]}$ and $g_{[-1]}$ when $L = \mathscr{P}(W)$, the power set of a certain set W.

(c) Determine $f^{[-1]}$ and $g_{[-1]}$ when $L = [-\infty, +\infty]^W$, the lattice of all functions on a set W with values in $[-\infty, +\infty]$.

(d) Try to say something on $f^{[-1]}$ and $g_{[-1]}$ when L is the complete lattice of all convex functions on $\mathbf{R}$ with values in $[-\infty, +\infty]$.

7.3. We know that $f_{[-1]} \leqslant f^{[-1]}$ if f is coincreasing. On the other hand $f^{[-1]} \leqslant f_{[-1]}$ when f is surjective. Find a lattice L and a function $f\colon L \to L$ such that $f^{[-1]}$ and $f_{[-1]}$ are not comparable (even with f strictly increasing and injective).

7.4. Prove that $\eta \mathbin{/_{\!\star}} \eta = \eta$ for all openings η, and hence, in view of Theorem 7.4.8, that $(f \mathbin{/_{\!\star}} f) \mathbin{/_{\!\star}} (f \mathbin{/_{\!\star}} f) = f \mathbin{/_{\!\star}} f$ for any mapping $f\colon L \to M$, where L is a set and M a complete lattice.

7.5. (a) Formulate and prove the statement dual to Theorem 7.4.8.

(b) Formulate and prove (at least in some special cases) the statement dual to *To be, or not to be, that is the question.*

7.6. Prove that if L is totally ordered, then every morphological filter is a strong filter.

7.7. Determine $f \circ f$, $f_{(1)}$, $f^{(1)}$, $f \circ f_{(1)}$, $f \circ f^{(1)}$ etc. in example 7.6.2.

8. Notions of topology

8.1. Mappings

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

$$\begin{aligned} f^*(B) &= \{x \in X;\, f(x) \in B\}, \qquad && B \subset Y; \\ f_*(A) &= \{f(x) \in Y;\, x \in A\}, && A \subset X. \end{aligned} \tag{8.1.1}$$

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

Given three sets X, Y and Z and a mapping $f\colon X \to Y$, it is customary to define a mapping $f^*\colon Z^Y \to Z^X$ by the formula $f^*(g) = g \circ f$, $g \in Z^Y$. The mapping $f^*(g)\colon X \to Z$ is called the *pull-back* of g. We thus have two mappings denoted by f^*, one mapping Z^Y into Z^X as just defined, and one mapping $\mathscr{P}(Y)$ into $\mathscr{P}(X)$ defined by (8.1.1). Hopefully this will not cause confusion, since the two definitions are compatible in a natural way if $Z = \mathbf{R}$: if $g \in \{0,1\}^Y \subset \mathbf{R}^Y$ is the characteristic function of a set $B \subset Y$, $g = \chi_B$, then $f^*(\chi_B) = \chi_{f^*(B)}$, where we use the notation from (8.1.1) in the right-hand side. Thus the pull-back of χ_B is the characteristic function of the set $f^*(B)$.

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

$$f^*(B_1 \cup B_2) = f^*(B_1) \cup f^*(B_2), \qquad f^*(B_1 \cap B_2) = f^*(B_1) \cap f^*(B_2), \tag{8.1.2}$$

$$f^*(B_1 \smallsetminus B_2) = f^*(B_1) \smallsetminus f^*(B_2), \text{ in particular } f^*\big(\complement B\big) = \complement f^*(B). \tag{8.1.3}$$

The formulas (8.1.2) can be generalized to infinite unions and intersections, implying that f^* is both a dilation and an erosion:

$$f^*\big(\bigcup B_j\big) = \bigcup f^*(B_j), \qquad f^*\big(\bigcap B_j\big) = \bigcap f^*(B_j).$$

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

$$f_*(A_1 \cup A_2) = f_*(A_1) \cup f_*(A_2) \text{ and } f_*\big(\bigcup A_j\big) = \bigcup f_*(A_j); \tag{8.1.4}$$

it is thus a dilation. But for intersections we have only $f_*(A_1 \cap A_2) \subset f_*(A_1) \cap f_*(A_2)$; f_* is in general not an erosion. Also there is in general no inclusion relation between $f_*\big(\complement A\big)$ and $\complement f_*(A)$.

We note that $A \subset f^*(B)$ if and only if $f_*(A) \subset B$, i.e., that $(\operatorname{hypo} f^*)^{-1} = \operatorname{epi} f_*$. By Corollary 7.3.8 this implies that $(f^*)^{[-1]} = f_*$ and that $(f_*)_{[-1]} = f^*$; f_* is the dilation corresponding to the erosion f^*. Moreover

$$f^* \circ f_* \circ f^* = f^* \text{ and } f_* \circ f^* \circ f_* = f_*.$$

It follows that $\big((f^*)^{[-1]}\big)_{[-1]} = f^* = \big((f^*)_{[-1]}\big)^{[-1]}$ and $\big((f_*)_{[-1]}\big)^{[-1]} = f_*$; cf. Proposition 7.3.9.

Proposition 7.3.3 shows that

$$f^* \circ f_* \geqslant \mathsf{Id}_{\mathscr{P}(X)} \text{ and } f_* \circ f^* \leqslant \mathsf{Id}_{\mathscr{P}(Y)}. \tag{8.1.5}$$

(These formulas can of course also be proved directly.) The first formula means that $f^*(f_*(A)) \supset A$; equality holds for all A if and only if f is injective. The second means that $f_*(f^*(B)) = B \cap \operatorname{im} f \subset B$; equality holds for all B if and only if f is surjective. Since f^* is also a dilation, Theorem 7.3.7 implies that $f^* \circ (f^*)_{[-1]} \leqslant \mathsf{Id}_{\mathscr{P}(X)}$. Proposition 7.3.3 implies that

$$(f_*)^{[-1]} \circ f_* \leqslant \mathsf{Id}_{\mathscr{P}(X)} \text{ and } (f^*)_{[-1]} \circ f^* \geqslant \mathsf{Id}_{\mathscr{P}(Y)}.$$

Finally we note that $(f^*)_{[-1]} = \mathsf{C} \circ f_* \circ \mathsf{C} = (f_*)^{\mathbf{d}} = \big((f^*)^{[-1]}\big)^{\mathbf{d}}$. (There seems to be no analogous formula for $(f_*)^{[-1]}$.)

We can now pass to mappings on a higher level by taking the upper and lower star again. There are four mappings on the next level:

(8.1.6)

$$(f^*)^*, (f_*)_*\colon \mathscr{P}(\mathscr{P}(X)) \to \mathscr{P}(\mathscr{P}(Y)) \text{ and } (f_*)^*, (f^*)_*\colon \mathscr{P}(\mathscr{P}(Y)) \to \mathscr{P}(\mathscr{P}(X));$$

two go in the same direction as f and two in the opposite direction. Of these four, two will be used to transport topologies, viz. $(f^*)^*$ and $(f^*)_*$.

8.2. Definition of topologies

A *topology* on a set X is a collection $\tau = \mathscr{U}(X)$ of subsets of X—thus an element of $\mathscr{P}(\mathscr{P}(X))$—which is closed under the formation of arbitrary unions and finite intersections. The elements of $\mathscr{U}(X)$ 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 $\emptyset$ and X are always open subsets of X.

In a metric space there is a topology defined by the metric: it consists of all unions of balls $B_<(c, r)$.

A topology can be given in several different ways. We define a set as *closed* if its complement is open. Then the family $\mathscr{F}(X)$ of all closed sets is closed under the formation of arbitrary intersections and finite unions. We may also impose these conditions as axioms, and then 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 $\mathscr{U}(X)$ 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 $\mathscr{V}(x)$, $x \in X$, of all neighborhoods of points in X satisfy the following conditions:

$$\text{If } V \in \mathscr{V}(x), \text{ then } x \in V; \tag{8.2.1}$$

$$\text{If } V \in \mathscr{V}(x) \text{ and } W \supset V, \text{ then } W \in \mathscr{V}(x); \tag{8.2.2}$$

$$\text{If } V, W \in \mathscr{V}(x), \text{ then } V \cap W \in \mathscr{V}(x); \tag{8.2.3}$$

(8.2.4) If $V \in \mathscr{V}(x)$, then there exists $W \in \mathscr{V}(x)$ such that $V \in \mathscr{V}(y)$ 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 $\mathscr{V}(x)$ for every $x \in X$ satisfying the axioms (8.2.1)—(8.2.4) and define a set U to be open if it belongs to $\mathscr{V}(x)$ 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 in $\mathscr{P}(X)$ 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 two 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 antiextensive, increasing, and idempotent. It is thus an opening in the algebraic sense, which of course means that it is a closure operator if we reverse the order: $A \leqslant B$ shall mean $A \supset B$. In addition to being an opening, 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. 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 $\mathscr{U}_1(X)$ and $\mathscr{U}_2(X)$ 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 $\mathscr{U}_1(X) \subset \mathscr{U}_2(X)$. Expressed in terms of closure operators, this means that $c_2 \leqslant c_1$ if c_j denotes the closure operator associated with $\mathscr{U}_j(X)$, $j = 1, 2$. The weakest topology is the *chaotic topology* $\{\varnothing, X\}$ and the strongest is the *discrete topology* $\mathscr{P}(X)$. 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 topologies.*[21] Of the four, only three are different in the sense that they cannot be obtained from another one by renaming the points.

8.3. 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 f^* is both a dilation and an erosion, it is clear that the family of all sets

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

is a topology. Here we have used the notation introduced in (8.1.1) at the next higher level: $f^*\colon \mathscr{P}(Y) \to \mathscr{P}(X)$, $(f^*)_*\colon \mathscr{P}(\mathscr{P}(Y)) \to \mathscr{P}(\mathscr{P}(X))$; see (8.1.6). For brevity

[21] Wacław Sierpiński, 1882—1969.

we shall denote $(f^*)_*(\tau) = (f^*)_*(\mathscr{U}(Y))$ by $f^{\leftarrow}(\tau) = f^{\leftarrow}(\mathscr{U}(Y))$, the *pull-back* of $\tau = \mathscr{U}(Y)$.

If $d\colon \mathscr{P}(Y) \to \mathscr{P}(Y)$ 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 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 $f^{\leftarrow}(\mathscr{U}(Y))$ 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 V 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 $\{f_*(A);\, A \in \mathscr{U}(X)\} \subset \mathscr{P}(Y)$. However, this family is usually not a topology on Y, since f_* is not an erosion. 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

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

the *push-forward* of the topology $\sigma = \mathscr{U}(X)$ on 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: $f^{\leftarrow}(f_{\rightarrow}(\mathscr{U}(X))) = \mathscr{U}(X)$. 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; $f_{\rightarrow}(f^{\leftarrow}(\mathscr{U}(Y))) = \mathscr{U}(Y)$. (This works so well because we did not use f_* but f^* in the definition of $f_{\rightarrow}$.)

However, if we have a closure operator c in X, we cannot define a closure operator in Y by anything like $c_{\rightarrow} = f_* \circ c \circ f^*$.

8.4. Continuous mappings

Let $f\colon X \to Y$ be a mapping of a topological space X into a topological space Y and let x be 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 can translate this well-known notion into the language of open sets, closed sets, and closure operators. One can prove that f is continuous if and only if $f^*(U) \in \mathscr{U}(X)$ for every $U \in \mathscr{U}(Y)$; in other words if and only if the topology $f^{\leftarrow}(\mathscr{U}(Y))$ is weaker than the topology $\mathscr{U}(X)$.

8.5. Connectedness

The family of all open and closed sets of a topological space X (sometimes called the *clopen* sets) is 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 of course only one such set.) A topological space is said to be *connected* if 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.[22] A *connectivity component* (sometimes called a "connected 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 8.5.1. *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 8.5.2. *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 $f^*(\{y\})$ 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.

Of the four topologies that can live on a space consisting of two points, only three are connected, and out of these, only two are different in the sense that they cannot be obtained from another one by renaming.

8.6. Quotient spaces

In particular we shall use Corollary 8.5.2 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. Now there exist very many surjective mappings $\mathbf{R} \to \mathbf{Z}$. It is not unnatural to restrict attention to increasing surjections $f\colon \mathbf{R} \to \mathbf{Z}$. Then $f^*(\{n\})$ is an interval for every integer n; denote its endpoints by a_n and $b_n \geqslant a_n$, so that

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

We can normalize the situation to $a_n = n - \frac{1}{2}$, $b_n = n + \frac{1}{2}$; this does not change the topology on $\mathbf{Z}$. Then $f(x) = \lfloor x + \frac{1}{2} \rfloor$ for all $x \in \mathbf{R} \smallsetminus (\mathbf{Z} + \frac{1}{2})$, and $f(n + \frac{1}{2}) = n$ or $f(n + \frac{1}{2}) = n + 1$ for $n \in \mathbf{Z}$. The topology is therefore determined if we know for which n we have $f(n + \frac{1}{2}) = n$. For every subset A of $\mathbf{Z}$ we get a topology on $\mathbf{Z}$ by declaring that $f(n+\frac{1}{2})$ shall be equal to n for $n \in A$ and that, for all other real numbers

[22] Hence the empty set is connected, which agrees with Bourbaki's definition (1961:I:§11:1). Adrien Douady (personal communication, 2000-06-26) claims that it would be more bourbakistic to declare the empty set not to be connected. However, I have decided to follow Bourbaki here.

x, we have $f(x) = \lfloor x + \frac{1}{2} \rfloor$. Thus A describes faithfully all topologies obtained from increasing surjections—the others are just too many . . .

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: $f(x) = \lfloor x + \frac{1}{2} \rfloor$ when $x \in \mathbf{R} \smallsetminus \left(\mathbf{Z} + \frac{1}{2}\right)$. So when $x = n + \frac{1}{2}$ we have a choice for each n: shall we define $f\left(n + \frac{1}{2}\right) = n$ or $f\left(n + \frac{1}{2}\right) = n + 1$? 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 8.5.2 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 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 $\left]\frac{1}{2}, \frac{3}{2}\right[$, 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. For the left topology, for instance, all subsets are connected.

8.7. Separation axioms

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 y will be denoted by $N(y)$. We note that $x \in N(y)$ if and only if $x \in \overline{\{y\}}$. The relation $x \in N(y)$ defines a preorder in X (see Definition 2.1.1). We shall denote it by $x \preccurlyeq y$; thus $x \preccurlyeq y$ if and only if $x \in N(y)$ if and only if $y \in \overline{\{x\}}$. It was introduced by Aleksandrov (1937:503). We shall call it the *specialization preorder* following Kong et al. (1991:905). (However, they defined it as the opposite preorder.)

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 (2.1.3)). It is quite reasonable to impose this axiom; if x belongs to $N(y)$ and vice versa, then x and y are indistinguishable from the point of view of topology: we cannot distinguish points from knowledge of the open sets to which they belong. We should therefore identify them and consider a quotient space.

So every topology τ on a set X defines a preorder $P(\tau) = \preccurlyeq$ in X, and this preorder is an order if and only if τ satisfies Kolmogorov's axiom.

Conversely, every preorder $\preccurlyeq$ in a set X defines a topology $T(\preccurlyeq)$ on X whose smallest neighborhoods are

$$N(y) = \{x \in X; x \preccurlyeq y\}.$$

This means that a subset A of X is open if and only if, for all $a \in A$ and all $y \in X$, $a \preccurlyeq y$ implies $y \in A$.

If we start with a preorder $\preccurlyeq$, define the topology $T(\preccurlyeq)$, and then define the preorder $P(T(\preccurlyeq))$, then we get back to the original preorder: $\preccurlyeq\, = P(T(\preccurlyeq))$; in other words $P \circ T$ is the identity on the set of all preorders. However, if we start with a topology

τ, construct the preorder $P(\tau)$, then the topology $T(P(\tau))$ defined by that preorder is not equal to τ in general; $T \circ P$ is not the identity. However, $T(P(\tau)) = \tau$ if and only if τ defines a smallest-neighborhood space as defined in the next section.

The separation axiom T_1 states that $N(x) = \{x\}$. It is too strong to be of interest for the spaces considered here. The specialization preorder in this case is the discrete order: 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 $y \in N(x)$. Hence two points are adjacent if and only if they are different and comparable for the specialization preorder.

8.8. 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 topological space with this property we shall call a *smallest-neighborhood space*. Another suitable name would be a *P. S. Aleksandrov space*, in honor of P. S. Aleksandrov,[23] who introduced them in his seminal paper (1935). It is equivalent to require that the union of an arbitrary family of closed sets be 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 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 (1937:504). He noted that the star of a set is a closure operation satisfying the two extra conditions $\overline{\varnothing} = \varnothing$ and $\overline{A \cup B} = \overline{A} \cup \overline{B}$ (see the discussion in section 2.2, in particular formula (2.2.6)), 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.

It is easy to see that a mapping $f \colon X \to Y$ between two smallest-neighborhood spaces is continuous if and only if it is increasing for the specialization preorder. Thus continuity in these spaces is actually order theoretic, and the smallest-neighborhood spaces are actually special cases of preordered sets. This means that the rich theory of (pre)ordered sets can be put to work here.

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 (1937:504) called *Eckpunkte* 'vertices', and the open points, i.e., the points x such that the singleton $\{x\}$ is open, he called *Grundpunkte* 'base points'.

[23] Pavel Sergeevič Aleksandrov (1896—1982); not to be confused with Aleksandr Danilovič Aleksandrov (1912—1999), for whom other spaces are named.

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 8.6. Thus in the Khalimsky topology the even points are *Eckpunkte* (vertices) and the odd points are *Grundpunkte* (base points) in Aleksandrov's terminology. In the specialization order, the vertices lie higher than the base points ($1 \preccurlyeq 0$, i.e., $1 \in N(0)$, and $1 \preccurlyeq 2$, i.e., $1 \in N(2)$).

A *Khalimsky interval* is an interval $[a,b]_{\mathbf{Z}} = [a,b]_{\mathbf{R}} \cap \mathbf{Z}$ 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 space* is the Cartesian product of n copies of $\mathbf{Z}$. Equivalently, we can define Khalimsky n-space 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 (Wyse 1970). The Khalimsky topology and the topology just defined are not comparable: none is stronger than the other. However, they are related, for if we rotate the Khalimsky plane by 45° and delete all points which are not open or closed, we obtain the other topology.

To exhibit some of the analogies between topological spaces and preordered sets, let us list some properties of continuous and increasing mappings.

Mappings $X \to Y$ between topological spaces	*Mappings $X \to Y$ between preordered sets*
X has the discrete topology $\Rightarrow$ all mappings are continuous	X has the discrete order $\Rightarrow$ all mappings are increasing
Y has the chaotic topology $\Rightarrow$ all mappings are continuous	Y has the chaotic preorder $\Rightarrow$ all mappings are increasing
X has the chaotic topology and Y has a Kolmogorov topology $\Rightarrow$ only the constants are continuous	X has the chaotic preorder and Y is ordered $\Rightarrow$ only the constants are increasing
Y has the discrete topology and X is connected $\Rightarrow$ only the constants are continuous	Y has the discrete order and X is connected $\Rightarrow$ only the constants are increasing

8.9. Exercises

8.1. Prove that if $f\colon X \to Y$ is a mapping of a set X into a set Y, then the closing $f^*/^\star f^*$ is equal to $f^* \circ f_* \geqslant \mathsf{Id}_{\mathscr{P}(X)}$. Similarly $f_*/\!\!_\star\, f_* = f_* \circ f^*$ (an opening; cf. Theorem 7.4.8).

8.2. Show that if we have a closure operator c in X and a mapping $f\colon X \to Y$, then we cannot define a closure operator in Y by something like $c_{\_\!\!\to} = f_* \circ c \circ f^*$. In general $c_{\_\!\!\to}$ will be neither extensive nor idempotent. Construct examples which show this. How shall we define the closure operator connected with the topology $f_{\rightarrow}(\mathscr{U}(X))$ on Y?

8.3. Express the continuity of a mapping $f\colon X \to Y$ from one topological space X into another, Y, in terms of the families of all closed sets $\mathscr{F}(X)$ and $\mathscr{F}(Y)$ as well as in terms of the closure operators in X and Y.

8.4. We have seen that a two-point space can have four topologies. How many topologies are there on a three-point space? How many of these are different in the sense that they cannot be obtained from another one by renaming the points? How many are connected? How many of these are different?

9. A closer look at the Khalimsky plane

9.1. Continuous functions

In this chapter we shall take a closer look at the Khalimsky plane, and in particular consider the Jordan curve theorem and Brouwer's fixed-point theorem in the new setting. We first explain the meaning of continuity.

We recall that a subset A of $\mathbf{Z}$ is open for the Khalimsky topology if and only if, for every even number $2n \in A$, also the two odd numbers $2n \pm 1$ belong to A. To construct the Khalimsky plane, we take the Cartesian product of two copies of the Khalimsky line. The topology is then determined by the rule that a subset A of the Khalimsky plane is open if and only if, for every pair of even numbers $x = (2m, 2n)$, all pairs $y \in \mathbf{Z}^2$ with $\|y - x\|_\infty \leqslant 1$ belong to A, for every pair $(2m, 2n+1) \in A$ also $(2m \pm 1, 2n+1) \in A$, and, finally, for every pair $(2m+1, 2n) \in A$, also $(2m+1, 2n \pm 1) \in A$.

A function $f\colon X \to Y$ from one smallest-neighborhood space into another is continuous at a point x if and only if the direct image of $N_X(x)$ is contained in $N_Y(f(x))$, or, equivalently, the inverse image of $N_Y(f(x))$ contains $N_X(x)$:

$$(9.1.1) \qquad f_*(N_X(x)) \subset N_Y(f(x)), \text{ equivalently } N_X(x) \subset f^*(N_Y(f(x))).$$

Here $N_X(x)$ and $N_Y(y)$ denote the smallest neighborhood of $x \in X$ and $y \in Y$, respectively. If we apply this to the case when $X = Y = \mathbf{Z}$, it means the following. If $x \in \mathbf{Z}$ is odd, the property always holds; if x is even and $f(x)$ is odd, it means that $f(x \pm 1) = f(x)$, and if x is even and also $f(x)$ is even, it means that $|f(x \pm 1) - f(x)| \leqslant 1$. In particular, a continuous function is Lip-1, but it must sometimes have intervals of constancy, viz. every time it takes an odd value at an even point (and hence also when it takes an even value at an odd point).

We observe that the following functions are continuous: (1) $\mathbf{Z} \ni x \mapsto a \in \mathbf{Z}$, where a is a constant; (2) $\mathbf{Z} \ni x \mapsto \pm x + c \in \mathbf{Z}$, where c is an even constant; (3) $\max(f, g)$ and $\min(f, g)$ if f, g are continuous. Actually every continuous function on a bounded Khalimsky interval can be obtained by a finite succession of the rules (1), (2), (3). Note that the function $x \mapsto x + 1$ is discontinuous.

For functions of two variables, $f\colon \mathbf{Z}^2 \to \mathbf{Z}$, (9.1.1) means the following. We list the eight possible parities of the triple $(x_1, x_2, f(x)) \in \mathbf{Z}^3$:

Parity of x_1	*Parity of* x_2	*Parity of* $f(x)$	*Condition*
Odd	Odd	Odd	None
Odd	Odd	Even	None
Odd	Even	Odd	$y_1 = x_1, \lvert y_2 - x_2\rvert \leqslant 1 \Rightarrow f(y) = f(x)$
Odd	Even	Even	$y_1 = x_1, \lvert y_2 - x_2\rvert \leqslant 1 \Rightarrow \lvert f(y) - f(x)\rvert \leqslant 1$
Even	Odd	Odd	$\lvert y_1 - x_1\rvert \leqslant 1, y_2 = x_2 \Rightarrow f(y) = f(x)$
Even	Odd	Even	$\lvert y_1 - x_1\rvert \leqslant 1, y_2 = x_2 \Rightarrow \lvert f(y) - f(x)\rvert \leqslant 1$
Even	Even	Odd	$\lVert y - x\rVert_\infty \leqslant 1 \Rightarrow f(y) = f(x)$
Even	Even	Even	$\lVert y - x\rVert_\infty \leqslant 1 \Rightarrow \lvert f(y) - f(x)\rvert \leqslant 1$

This table lists what it means for f to be continuous at a particular point x. However, the description becomes much simpler if we use the specialization order $\preccurlyeq$, for then a continuous function is just an increasing function. We know that, in $\mathbf{Z}$,

$$\cdots \preccurlyeq -2 \succcurlyeq -1 \preccurlyeq 0 \succcurlyeq 1 \preccurlyeq 2 \succcurlyeq 3 \preccurlyeq 4 \succcurlyeq \cdots$$

In $\mathbf{Z}^2$, $(0,0) \succcurlyeq (1,0), (0,1) \succcurlyeq (1,1)$ and, in general,

$$(2m, 2n) \succcurlyeq (2m+1, 2n), (2m, 2n+1) \succcurlyeq (2m+1, 2n+1) \text{ for all } m, n \in \mathbf{Z}.$$

So continuity at x boils down to $x \preccurlyeq y \Rightarrow f(x) \preccurlyeq f(y)$ for all $y \in \mathbf{Z}^2$; continuity everywhere to the same implication but now for all $x, y \in \mathbf{Z}^2$. For example, if both components of x are odd, the only y which satisfies $x \preccurlyeq y$ is $y = x$, so $f(x) \preccurlyeq f(y)$ holds automatically. If, on the other hand both components of x are even, then $\{y; x \preccurlyeq y\} = B_{\leqslant}(x, 1)$ for the l^∞ norm, and if $f(x)$ in addition is odd, then $f(x) \preccurlyeq f(y)$ holds only for $f(y) = f(x)$, so f must be constant on $B_{\leqslant}(x, 1)$.

We note that if $x, y \in \mathbf{Z}$ and $x \preccurlyeq y$, then $|x - y| \leqslant 1$. Conversely, if $|x-y| \leqslant 1$, then either $x \preccurlyeq y$ or $y \preccurlyeq x$. Hence $|x - y| \leqslant 1$ implies $|f(x) - f(y)| \leqslant 1$ for any continuous function $f \colon \mathbf{Z} \to \mathbf{Z}$, and we see that f is Lip-1. In two variables we have the same conclusion. In the proof of this fact we shall need the following notation. For any two points $x, y \in \mathbf{Z}^2$ we define $q(x, y) = (x_1, y_2)$. The four points $x, y, q(x, y), q(y, x)$ thus form a rectangle (perhaps degenerate); if $y_j = x_j \pm 1$, $j = 1, 2$, they form a square with side length 1.

Theorem 9.1.1. *A continuous function* $f \colon \mathbf{Z}^2 \to \mathbf{Z}$ *is Lip-*1 *for the* l^∞ *norm. More generally, the conclusion holds for any continuous function* $f \colon X \to \mathbf{Z}$*, where* X *is a connected subset of* $\mathbf{Z}^2$ *such that* $q(x, y), q(y, x) \in X$ *for all* $x, y \in X$ *such that* $y_j = x_j \pm 1$, $j = 1, 2$*, and we do not have* $x \preccurlyeq y$*, nor* $y \preccurlyeq x$.

Proof. It is enough to prove that if $\lVert x - y\rVert_\infty \leqslant 1$, then $|f(x) - f(y)| \leqslant 1$. If $x, y \in X$ with $x \preccurlyeq y$, then $f(x) \preccurlyeq f(y)$, which in turn implies that $|f(x) - f(y)| \leqslant 1$. Assume now that $\lVert x - y\rVert_\infty \leqslant 1$ but that we do not have $x \preccurlyeq y$, nor $y \preccurlyeq x$. Then we have essentially the case $x = (1, 0)$, $y = (0, 1)$. We have $q(y, x) = (0, 0) \succcurlyeq x, y \succcurlyeq (1, 1) = q(x, y)$, which implies that $f(0, 0) \succcurlyeq f(x), f(y) \succcurlyeq f(1, 1)$. (By hypothesis both $(0, 0)$ and $(1, 1)$ belong to X.) Hence $|f(0, 0) - f(1, 1)| \leqslant 1$. Also, $|f(0, 0) - f(x)| \leqslant 1$ and $|f(x) - f(1, 1)| \leqslant 1$, which implies that $f(x) = f(0, 0)$ or $f(1, 1)$. Similarly, $f(y) = f(0, 0)$ or $f(y) = f(1, 1)$. Therefore $|f(x) - f(y)| \leqslant 1$.

This result holds for many subsets X of $\mathbf{Z}^2$, but not for all:

Example 9.1.2. Let $X = \{x \in \mathbf{Z}^2; \|x\|_1 \leqslant 1\}$, the l^1 ball in $\mathbf{Z}^2$ and define $f(x) = x_1 - x_2$. Then X is connected for the Khalimsky topology and f is continuous, but f is not Lip-1. We note that $q(y,x) = (0,0) \in X$ but that $q(x,y) = (1,1) \notin X$ if we take $x = (1,0) \in X$, $y = (0,1) \in X$. So both $q(x,y)$ and $q(y,x)$ need to be elements of X.

Theorem 9.1.3. *A function $f\colon \mathbf{Z}^2 \to \mathbf{Z}$ is continuous if and only if it is separately continuous. More generally, the equivalence holds for any function $f\colon X \to \mathbf{Z}$ where X is a subset of $\mathbf{Z}^2$ such that one of $q(x,y)$, $q(y,x)$ belongs to X if $y_j = x_j \pm 1$ and $x \preccurlyeq y$.*

Proof. Assume that f is separately continuous and that $x \preccurlyeq y$. Then we shall prove that $f(x) \preccurlyeq f(y)$. If $x_1 = y_1$, then $x_2 \preccurlyeq y_2$, and the inequality $f(x) \preccurlyeq f(y)$ follows from the separate continuity of the function $x_2 \mapsto f(x)$ for a fixed x_1. The conclusion is similar if $x_2 = y_2$; then the continuity of $x_1 \mapsto f(x)$ for a fixed x_2 does the job. The case when $x_1 \neq y_1$ and $x_2 \neq y_2$ remains to be considered. Then $y_j = x_j \pm 1$. One of the points $q(x,y)$ and $q(y,x)$ belongs to X; let z be one of them that does. Then clearly $x \preccurlyeq z \preccurlyeq y$, which in view of the separate continuity implies $f(x) \preccurlyeq f(z)$ and $f(z) \preccurlyeq f(y)$, and we are done.

Example 9.1.4. Let $X = \{0\} \cup \{x \in \mathbf{Z}^2; |x_1| = |x_2| = 1\}$. This set is connected for the Khalimsky topology. Every function $f\colon X \to \mathbf{Z}$ is separately continuous, but not all of them are continuous. With $x = (0,0)$ and $y = (1,1)$ we see that none of $q(x,y) = (0,1)$ and $q(y,x) = (1,0)$ belongs to X.

Like in real analysis there is an intermediate-value theorem for the Khalimsky line:

Theorem 9.1.5. *Let two continuous functions $f, g\colon I \to \mathbf{Z}$ be given on a Khalimsky interval $I = [a,b]_{\mathbf{Z}}$. Assume that there are points $s, t \in I$ with $f(s) \geqslant g(s)$ and $f(t) \leqslant g(t)$. Then there exists a point p, intermediate between s and t, such that $f(p) = g(p)$.*

Proof. Without loss of generality we may assume that $s \leqslant t$. Define

$$M = \{x \in \mathbf{Z}; s \leqslant x \leqslant t \text{ and } f(x) \geqslant g(x)\}.$$

Clearly $s \in M$, so M is not empty. Let $p = \max_{x \in M} x$. If $p = t$, then $f(t) = g(t)$ and we are done; if not, $p + 1 \leqslant t$. Then we must have $f(p) \geqslant g(p)$, $f(p+1) < g(p+1)$. We claim that $f(p) = g(p)$. If this were not true, we would have $f(p) \geqslant g(p) + 1$, $f(p+1) \leqslant g(p+1) - 1$. Because of the Lipschitz continuity, the only possibility then would be $f(p) = g(p) + 1$ and $f(p+1) = g(p+1) - 1$. But even this situation is impossible. If p and $f(p)$ are of different parity, then $f(p+1) = f(p)$, which would require a jump of two units in g. If on the other hand p and $f(p)$ are of the same parity, then p and $g(p)$ are of different parity, so that $g(p+1) = g(p)$, requiring a jump of two units in f. This contradiction leaves us with the only possibility $f(p) = g(p)$.

9.2. A fixed-point theorem

Let us say that a topological space X has the *fixed-point property* if every continuous mapping $f\colon X \to X$ possesses a *fixed point*, i.e., a point p such that $f(p) = p$. The famous Brouwer[24] fixed-point theorem states that a closed ball in $\mathbf{R}^n$ has the fixed-point property. This theorem is a deep result when $n \geqslant 2$ but easy to prove if $n = 1$: we may take

$$p = \sup_x (x; f(x) \geqslant x). \tag{9.2.1}$$

If $f\colon [0,1] \to [0,1]$ is a mapping from a compact interval into itself which is increasing for the usual order inherited from the real axis, then it also has a fixed point, and formula (9.2.1) again yields a fixed point. This generalizes to complete lattices; if $f\colon L \to L$ is an increasing mapping of a complete lattice L into itself, then the point p defined by (9.2.1) is a fixed point of f. This is Tarski's fixed point theorem (1955), exhibiting one of the many analogies between continuous mappings and increasing mappings. However, a Khalimsky interval of more than two points is not a lattice.

In this section we shall prove that certain subsets of the Khalimsky plane possess the fixed-point property. Since the Khalimsky line and the Khalimsky plane are ordered sets and the continuous mappings are precisely the increasing mappings, there are theorems from the theory of ordered sets that can be applied here; cf. Baclawski & Björner (1979). However, the proofs presented here are more direct in the context of digital geometry.

A topological space cannot have the fixed-point property unless it is connected. Indeed, if U and V are two disjoint nonempty open sets whose union is the whole space X, then we may define $f(x) = b$ for all $x \in U$ and $f(x) = a$ for all $x \in V$, where a is an arbitrary point in U and $b \in V$. Then all inverse images are open, for they are either empty, equal to U, equal to V, or equal to X. Thus f is certainly continuous, but it has no fixed point.

On the other hand, connectedness is by no means sufficient. The mapping $f(x) = -x$ is continuous in $\mathbf{R}^n$, and in fact in any abelian group with a reasonable topology; it has no fixed point in X if $0 \notin X$. Any nonempty subset X of such a group with $0 \notin X$ and such that $x \in X$ implies $-x \in X$ will be a space such that some continuous selfmapping is without fixed point.

In a finite set X with N points there are N^N selfmappings $X \to X$. Out of these, $(N-1)^N$ do not have fixed points; there are thus $N^N - (N-1)^N$ mappings which have a fixed point. The proportion of mappings with a fixed point is $1 - (1 - 1/N)^N$, which tends to $1 - 1/e$ (about 63 percent) when N tends to infinity. If we introduce a topology on X we may ask how many of the N^N mappings are continuous. Let us denote by C the number of continuous mappings. If the space possesses the fixed-point property the number of mappings of different kinds can then be listed as follows. The table contains only one unknown, C, the number of continuous mappings.

[24] Luitzen Egbertus Jan Brouwer, 1881—1966.

	Continuous	*Discontinuous*	*Sum*
Fixed point	C	$N^N - (N-1)^N - C$	$N^N - (N-1)^N$
No fixed point	0	$(N-1)^N$	$(N-1)^N$
Sum	C	$N^N - C$	N^N

We shall now prove that a continuous mapping of an interval into itself has a fixed point. Later we shall consider subsets of the Khalimsky plane, which we shall attack using an implicit-function theorem.

Theorem 9.2.1. *Let us define a subset $C_\#(\mathbf{Z}, \mathbf{Z})$ of the set of all continuous mappings of $\mathbf{Z}$ into itself,*

$$C_\#(\mathbf{Z}, \mathbf{Z}) = \{f \in C(\mathbf{Z}, \mathbf{Z}); \exists s \in \mathbf{Z}, f(s) \geqslant s \text{ and } \exists t \in \mathbf{Z}, f(t) \leqslant t\}. \tag{9.2.2}$$

Then $f \in C(\mathbf{Z}, \mathbf{Z})$ has a fixed point if and only if $f \in C_\#(\mathbf{Z}, \mathbf{Z})$.

Proof. If f has a fixed point, it is obvious that f belongs to $C_\#(\mathbf{Z}, \mathbf{Z})$. The converse is a special case of the intermediate-value theorem (Theorem 9.1.5), taking g as the identity.

Corollary 9.2.2. *Every bounded Khalimsky interval has the fixed-point property.*

Proof. Let $f\colon I \to I$ be a continuous mapping, where $I = [a, b]_\mathbf{Z}$ is a bounded interval. Extend f to a mapping $g\colon \mathbf{Z} \to \mathbf{Z}$ by defining $g(x) = f(a)$ for $x < a$ and $g(x) = f(b)$ for $x > b$. Then it is easy to check that g is continuous, and it is obvious that g belongs to $C_\#(\mathbf{Z}, \mathbf{Z})$. Thus it has a fixed point $p \in \mathbf{Z}$, but as $p \in \operatorname{im} g \subset I$, p is a fixed point also of f.

Example 9.2.3. For a Khalimsky interval $\{a, a+1\}$ consisting of two points only, there are four mappings: the two constant mappings, the identity, and the one interchanging a and $a+1$. The first three obviously have a fixed point; the fourth does not. But it is discontinuous. Thus the statistics looks like this ($N = 2$, $C = 3$):

	Continuous	*Discontinuous*	*Sum*
Fixed point	3	0	3
No fixed point	0	1	1
Sum	3	1	4

Theorem 9.2.4. *Every bounded Khalimsky rectangle $I \times J \subset \mathbf{Z}^2$ has the fixed-point property.*

We shall prove this result using an implicit-function theorem. In real analysis there is an implicit-function theorem which says the following. If f is a real-valued function of class C^1 defined in an open subset Ω of $\mathbf{R}^2$ with a zero $a \in \Omega$, and its derivative with respect to the first variable, $\partial f/\partial x_1$, is non-zero at a, then there is a function g of one variable defined near a_2 such that $f(g(x_2), x_2) = 0$ for all x_2 near a_2. This function

g is also of class C^1. One says that the equation $f(x_1, x_2) = 0$ defines x_1 *implicitly* as a function of x_2; the equation $f(x_1, x_2) = 0$ is equivalent to the *explicit* formula $x_1 = g(x_2)$ near a. The result is local in the sense that we only assert something in a neighborhood of a.

In the digital plane we have an implicit-function theorem of a similar kind. We shall solve the implicit equation $f(x, y) = x$. (In real analysis, this is equivalent to $f(x, y) - x = 0$, but the left-hand side $f(x, y) - x$ is not continuous in general in the digital case even if f is.)

Theorem 9.2.5. *Let us define an order, and hence a topology, in the space* $C(\mathbf{Z}, \mathbf{Z})$ *of continuous mappings* $\mathbf{Z} \to \mathbf{Z}$ *by declaring that* $f \preccurlyeq g$ *if and only if* $f(x) \preccurlyeq g(x)$ *for all* $x \in \mathbf{Z}$. *Let us also define a mapping* $F\colon C(\mathbf{Z}, \mathbf{Z}) \to \mathbf{Z} \cup \{-\infty, +\infty\}$ *by*

$$F(g) = \sup_{x \in \mathbf{Z}}(x; g(x) \geqslant x), \qquad g \in C(\mathbf{Z}, \mathbf{Z}). \tag{9.2.3}$$

The mapping F *is continuous in* $C_{\#}(\mathbf{Z}, \mathbf{Z})$ *where it is finite.*

If the set of fixed points of a mapping $g \in C_{\#}(\mathbf{Z}, \mathbf{Z})$ is bounded to the right, then $F(g)$ is its largest fixed point; otherwise $F(g) = +\infty$.

Lemma 9.2.6. *The mapping* F *defined by (9.2.3) satisfies*

$$g(x) \geqslant x \text{ if and only if } F(g) \geqslant x, \qquad x \in \mathbf{Z}, g \in C_{\#}(\mathbf{Z}, \mathbf{Z}). \tag{9.2.4}$$

Proof. The implication to the right follows easily from the definition of F. For the other implication we use the Lip-1 property of g: if $g(x) < x$, then $g(x') < x'$ for all $x' \geqslant x$, which implies that $F(g) < x$.

Proof of Theorem 9.2.5. We shall prove that $f \preccurlyeq g$ implies $F(f) \preccurlyeq F(g)$. Fix two functions f and g satisfying $f \preccurlyeq g$. We consider the two cases $F(f)$ odd, $F(f)$ even.

Case 1. Assume that $F(f)$ is odd. Without loss of generality we may assume that $F(f) = 1$, thus that $f(1) = 1$ and that $f(2) = 1$ or $f(2) = 0$. We shall prove that $F(g)$ is 0, 1, or 2. We write the inequality $0 \leqslant F(g) \leqslant 2$ as $0 \leqslant F(g) < 3$ and then translate it using Lemma 9.2.6 as $0 \leqslant g(0)$ and $g(3) < 3$. We know that $g(1) \succcurlyeq f(1) = 1$, which implies that $g(1) \geqslant 0$ and that $g(0) \geqslant 0$. Hence it only remains to be proved that $g(3) < 3$.

Case 1.1. Assume that $f(2) = 1$. Then $f(1) = f(2) = f(3) = 1$, so that $g(3) \succcurlyeq f(3) = 1$ and $g(3) \leqslant 2 < 3$.

Case 1.2. Assume that $f(2) = 0$. Then $g(2) \succcurlyeq f(2) = 0$, which implies that $g(2) = 0$; by Lipschitz continuity $g(3) \leqslant 1 < 3$.

Case 2. If $F(f)$ is even, we may assume that $F(f) = 0$ without loss of generality. We shall then prove that $F(g) = 0$. We write this as $0 \leqslant F(g) < 1$ and translate it using Lemma 9.2.6 as $0 \leqslant g(0)$ and $g(1) < 1$. From $F(f) = 0$ it follows that $f(0) = 0$ and that $f(1) = 0$ or $f(1) = -1$. Since $0 = f(0) \preccurlyeq g(0)$, it follows that $g(0) = 0$. It remains to be proved that $g(1) < 1$.

Case 2.1. Assume that $f(1) = -1$. Then $g(1) \succcurlyeq f(1) = -1$, which implies that $g(1) \leqslant 0 < 1$.

Case 2.2. Assume finally that $f(1) = 0$. Then $g(1) \succcurlyeq f(1) = 0$, so $g(1) = 0 < 1$.

Corollary 9.2.7. *Let $f\colon I \times J \to I$ be a continuous function defined in a rectangle $I \times J$, where I and J are Khalimsky intervals, I being bounded. Then*

$$h(y) = \max_{x\in I}(x; f(x,y) \geqslant x), \qquad y \in J, \tag{9.2.5}$$

is continuous in J. The function h satisfies $f(h(y),y) = h(y)$.

Proof. The function h is given by $h(y) = F(x \mapsto f(x,y))$. It depends continuously on the function $x \mapsto f(x,y)$, which in turn is a continuous function of y.

Example 9.2.8. It is natural to ask if the function

$$h(y) = \max_{x\in I}\bigl(x; f(x,y) \geqslant g(x,y)\bigr), \qquad y \in J,$$

is continuous for more general choices of functions f and g. That this is not always so is shown by the example $f(x,y) = y$, $g(x,y) = \min(x,0)$. Then $h(y) = y$ if $y < 0$ and $h(y) = b$ if $y \geqslant 0$, assuming the interval I to be $[a,b]_{\mathbf{Z}}$. This yields a discontinuous function if $a \leqslant -1$, $b \geqslant 1$.

Proof of Theorem 9.2.4. The mapping f has two components f_1 and f_2. As in (9.2.5) we let $h(y)$ denote the largest fixed point of the partial mapping $x \mapsto f_1(x,y)$ for a fixed y, $h(y) = F(x \mapsto f_1(x,y))$. So $f_1(h(y),y) = h(y)$ for all $y \in J$. We then form the composition $k(y) = f_2(h(y),y)$. As a composition of continuous mappings it is continuous, and by the one-dimensional theorem it possesses a fixed point q, $k(q) = q$. Collecting what we have, we see that $f_1(h(q),q) = h(q)$ and that $k(q) = f_2(h(q),q) = q$, which means that we have proved that $f(h(q),q) = (f_1(h(q),q), f_2(h(q),q)) = (h(q),q)$, thus that $(h(q),q)$ is a fixed point.

Example 9.2.9. The Khalimsky square $\{0,1\}^2 \subset \mathbf{Z}^2$. There are $N^N = 4^4 = 256$ selfmappings of $\{0,1\}^2$, of which $(N-1)^N = 3^4 = 81$ do not have fixed points. The proportion of mappings without a fixed point is $81/256 \approx 0.3164$, as expected slightly lower than the limit $1/e \approx 0.3679$. The remaining $256 - 81 = 175$ have a fixed point. Of the 16 mappings $\{0,1\}^2 \to \{0,1\}$, 6 are continuous. There are therefore $6^2 = 36$ continuous mappings $\{0,1\}^2 \to \{0,1\}^2$, and we know already that they all have fixed points.

The table of different kinds of mappings therefore looks as follows ($N = 4$, $C = 6^2$).

	Continuous	*Discontinuous*	*Sum*
Fixed point	36	139	175
No fixed point	0	81	81
Sum	36	220	256

In this simple case we can of course see directly that all continuous mappings have a fixed point. Indeed, of the 6 continuous mappings $\{0,1\}^2 \to \{0,1\}$, five map $(0,0)$ to 0; the remaining one is the constant 1. Therefore, of the 36 continuous mappings

$\{0,1\}^2 \to \{0,1\}^2$, $(0,0)$ is a fixed point except when one of the components is the constant 1. Thus they all have a fixed point.

It is easy to generalize the implicit-function theorem and the fixed-point theorem to somewhat more general sets. We formulate an example of the latter:

Theorem 9.2.10. *Let X be a subset of $\mathbf{Z}^2$ defined as*

$$X = \{(x,y) \in \mathbf{Z} \times J; \varphi(y) \leqslant x \leqslant \psi(y)\},$$

where J is a bounded Khalimsky interval, and φ and ψ two continuous functions defined on J. Assume that $\varphi(y) < \psi(y)$ for all $y \in J$. Then X has the fixed-point property.

Proof. Take an interval $I = [a,b]_{\mathbf{Z}}$ which is so large that $\varphi(y), \psi(y) \in I$ for all $y \in J$. Then extend $f\colon X \to X$ to a mapping $g\colon I \times J \to I \times J$ by defining $g(x,y) = \varphi(y)$ when $a \leqslant x < \varphi(y)$ and $g(x,y) = \psi(y)$ when $\psi(y) < x \leqslant b$. Then g is continuous. It must have a fixed point in $I \times J$ according to Theorem 9.2.5. However, the fixed point must actually lie in X and be a fixed point of f.

The result on separate continuity (Theorem 9.1.3) makes it easy to go up in dimension.

9.3. Jordan curve theorems

There is a Jordan curve theorem in the Khalimsky plane. During the lectures I discussed this; here [at least in this version of the notes] I just refer to (Kiselman 2000) and the references therein.

9.4. Exercises

9.1. How many continuous selfmappings are there in a Khalimsky interval with three elements?

9.2. How many continuous selfmappings are there in a Khalimsky interval with four elements like $\{0,1,2,3\}$?

9.3. Prove in detail that there are six continuous mappings $\{0,1\}^2 \to \{0,1\}$ and hence that there are 36 continuous mappings $\{0,1\}^2 \to \{0,1\}^2$. Compare with the previous exercise; both spaces have four elements.

9.4. Try to estimate the number C of continuous mappings $I \to J$ between Khalimsky intervals, and more generally mappings $I_1 \times \cdots \times I_m \to J_1 \times \cdots \times J_m$ between boxes in Khalimsky spaces. Compare with the number of Lip-1 mappings.

10. Digitization

10.1. What is a digitization?

Digital geometry is about geometry in digital spaces—however, we shall not at this point give a formal definition. Suffice it to say that *digital*[25] is used here as opposed

[25] The word itself comes the Latin *digitus*, meaning ‘finger, toe’ and being related to the Greek *daktylos* with the same meaning. The European plant foxglove has received its scientific name *Digitalis purpurea* because of its finger-like corollas. The Greek word is also the origin of the name of a fruit, *date*. If you look at dates, not in a box, but growing high up in a palm *Phoenix dactylifera*, you will appreciate the similarity—*dactylifera* means ‘carrying fingers.’ So digital geometry is about counting on fingers and toes (perhaps implying using a system of base twenty) but if that seems to you not to be sweet enough, think of counting dates.

to *continuous.* The space $\mathbf{R}^n$ is a space where we do continuous geometry; the space $\mathbf{Z}^n$ is an example of a space where we do digital geometry. However, we shall take a more general approach.

In this chapter we shall first discuss what a good digitization should mean, and then study the notion of a digital line.

Let X be a set and Z an arbitary subset of X. (Think of X as $\mathbf{R}^2$ and Z as $\mathbf{Z}^2$ if you like.) If we want to digitize X we may start with a mapping $f\colon X \to Z$ and then define the digitization of a set A as $f_*(A) = \{f(x); x \in A\}$; cf. (8.1.1).

However, this approach is too narrow; it is often not possible to start with a pointwise mapping. Instead, we shall define here a digitization of X into Z as a mapping $F\colon \mathscr{P}(X) \to \mathscr{P}(Z)$ with certain desirable properites. We shall think of $F(A)$ as a digital representation of A. A very simple such representation is $F(A) = A \cap Z$, but it is not very faithful, since many sets are mapped to the empty set, for example $A = X \smallsetminus Z$. (However, it works for sufficiently fat sets.) One desirable condition is therefore that $F(A)$ be empty only if A is empty. We also remark that the mapping $F(A) = A \cap Z$ is not of the form $F = f_*$ if $Z \neq X$.

We recall that the mappings f_* are dilations; see (8.1.4). It seems desirable to require in general that a digitization $F\colon \mathscr{P}(X) \to \mathscr{P}(Z)$ be a dilation. In particular this means that it is determined by its images on points, i.e., $F(A) = \bigcup_{x\in A} F(\{x\})$. So it is enough to know the digitization of an arbitrary point in X; however, nothing requires the $F(\{x\})$ to be singleton sets.

The following setup seems to be sufficiently flexible.

Definition 10.1.1. *Let two sets X and Z be given, Z being a subset of X. Let there be given, for every $p \in Z$, a subset $C(p)$ of X, called the* cell with nucleus p. *Then the* digitization *determined by these cells is the mapping $F\colon \mathscr{P}(X) \to \mathscr{P}(Z)$ defined by*

$$F(\{x\}) = \{p \in Z; x \in C(p)\}, \qquad x \in X, \tag{10.1.1}$$

and

$$F(A) = \bigcup_{x\in A} F(\{x\}) = \{p \in Z; A \text{ meets } C(p)\}, \qquad A \in \mathscr{P}(X). \tag{10.1.2}$$

We may think of the cell $C(p)$ as a pixel or voxel, and of p as its address. If we think of C as a mapping $C\colon Z \to \mathscr{P}(X)$, then $F(A) = C^*(\mathscr{C}_A)$, where $C^*\colon \mathscr{P}(\mathscr{P}(X)) \to \mathscr{P}(Z)$ is defined by (8.1.1) and $\mathscr{C}_A \in \mathscr{P}(\mathscr{P}(X))$ is the family of all cells which meet A.

It is clear that a digitization in this sense is always a dilation. As already pointed out, it is desirable that a nonempty set have a nonempty digitization; this is true if and only if the union of all cells is equal to the whole space X.

If X is an abelian group and Z a subgroup, it is desirable that the digitization commute with translations, which means that $C(p) = C(0) + p$ for all $p \in Z$. Indeed, if $C(p) - p$ varies too much, it is easy to construct strange examples.

Example 10.1.2. A simple choice is $C(p) = \{p\}$. This yields the digitization $F(A) = Z \cap A$ already mentioned. If the set is fat, this digitization may work out well. In an abelian group with a metric we can even fatten the set using a dilation, defining $C(p)$ to be a ball $B_\leqslant(p, r)$ of radius r; this yields $F(A) = Z \cap (A + B_\leqslant(0, r))$.

Example 10.1.3. If $X = \mathbf{R}$ and $Z = \mathbf{Z}$ we may choose $C(p) = \left[p - \frac{1}{2}, p + \frac{1}{2}\right]$. Then every set has a nonempty digitization, but the half-integers have a digitization consisting of two points. If we choose instead $C(p) = \left]p - \frac{1}{2}, p + \frac{1}{2}\right[$, then the digitization of a half-integer is empty. As a compromise we may choose $C(p) = \left]p - \frac{1}{2}, p + \frac{1}{2}\right]$; the digitization of a point is then always a point: $F(\{x\}) = \left\{\left\lceil x - \frac{1}{2}\right\rceil\right\}$. But then a new disadvantage appears: this digitization does not commute with the reflection $x \mapsto -x$.

Example 10.1.4. If $X = \mathbf{R}^2$ and $Z = \mathbf{Z}^2$ we may construct digitizations from what we have already done on the real axis. We may take

$$C(p) = \left[p_1 - \tfrac{1}{2}, p_1 + \tfrac{1}{2}\right] \times \left[p_2 - \tfrac{1}{2}, p_2 + \tfrac{1}{2}\right], \qquad p \in \mathbf{Z}^2,$$

and similarly for the open and half-open intervals. Another choice is not to take the Cartesian product but to define the cell with nucleus p as

$$\begin{aligned} C_R(p) = &\left\{x; x_1 = p_1 \text{ and } p_2 - \tfrac{1}{2} < x_2 \leqslant p_2 + \tfrac{1}{2}\right\} \\ &\cup \left\{x; p_1 - \tfrac{1}{2} < x_1 \leqslant p_1 + \tfrac{1}{2} \text{ and } x_2 = p_2\right\}. \end{aligned} \tag{10.1.3}$$

Thus $C_R(p)$ is a cross with center at p. This is the digitization used by Rosenfeld (1974). It is based on the mapping $\mathbf{R} \ni x \mapsto \left\lceil x - \frac{1}{2}\right\rceil \in \mathbf{Z}$, already mentioned, a digitization of $\mathbf{R}$ which takes a non-half-integer to the closest integer and moves down by one half in the case of half-integers. Let us call it the *Rosenfeld digitization* of $\mathbf{R}^2$.

It is clear that in this case the union of the cells is very small compared with $\mathbf{R}^2$, so that many sets have empty digitization. However, the union of all cells is equal to all grid lines $(\mathbf{R} \times \mathbf{Z}) \cup (\mathbf{Z} \times \mathbf{R})$, so that every straight line has a nonempty digitization. The same is true of a sufficiently long straight line segment. Thus this digitization can be used in the study of digital straight lines. Note that the family of all cells is disjoint, which implies that the digitization of a point is either empty or a singleton set.

The definition as such says nothing about how close a digitization of a point is to the point. To achieve this we must of course add some requirement that points in the cell $C(p)$ shall be reasonably close to p. This leads us to the next topic, that of Voronoi cells.

10.2. Voronoi cells

Let a metric space X be given as well as a subset Z. The metric of X shall be denoted by d. For a point $x \in X$ we view the points in Z close to x as approximants; there may be a best approximant. Given $p \in Z$ we shall give a name to the set of all x for which this particular p is a (not necessarily unique) best approximant: the *Voronoi cell*[26] *with nucleus* p is

$$\mathsf{Vo}(p) = \{x \in X; \forall q \in Z, d(x,p) \leqslant d(x,q)\}, \qquad p \in Z. \tag{10.2.1}$$

Thus $x \in \mathsf{Vo}(p)$ if and only if p is a best approximant of x. We also define the *strict Voronoi cell* as

$$\mathsf{Vo}_\mathsf{s}(p) = \{x \in X; \forall q \in Z \smallsetminus \{p\}, d(x,p) < d(x,q)\} \qquad p \in Z. \tag{10.2.2}$$

[26] Named for Georgi Fedoseevič Voronoi (1868—1908)....

Thus $x \in \mathsf{Vo}_{\mathbf{s}}(p)$ if and only if p is the unique best approximant of x. Finally, one might define the *very strict Voronoi cell* as

$$\mathsf{Vo}_{\mathbf{vs}}(p) = \{x \in X; d(x,p) < \inf_{q \in Z \smallsetminus \{p\}} d(x,q)\}, \qquad p \in Z. \tag{10.2.3}$$

It is easy to construct examples where the very strict Voronoi cell is different from the strict Voronoi cell, but in all applications we are interested in they are equal.

Two different strict Voronoi cells are disjoint. Even more can be said: a (nonstrict) Voronoi cell is disjoint from every strict Voronoi cell with a different nucleus. The union of all strict Voronoi cells is almost equal to the whole space X; there is only some garbage left out: these are the points which have at least two best approximants in Z. However, since we are mathematicians, we do not have the right to throw away that garbage; we must be careful and consider both the strict and the nonstrict Voronoi cells.

We now return to the topic of digitization. It seems reasonable that the digitization of a point should be contained in the set of all nuclei of Voronoi cells which contain that point. After all, these nuclei are the best approximants in Z of the point. This argument leads us to the following definition.

Definition 10.2.1. *Let X be a metric space and Z a subset of X such that $Z \cap B_<(c,r)$ is finite for all $c \in X$ and all $r \in \mathbf{R}$. A* Voronoi digitization of X into Z *is a dilation* $\mathsf{Dig}\colon \mathscr{P}(X) \to \mathscr{P}(Z)$ *such that*

$$\mathsf{Dig}(\{x\}) \subset \{p \in Z; x \in \mathsf{Vo}(p)\}. \tag{10.2.4}$$

Note that if x belongs to some strict Voronoi cell $\mathsf{Vo}_{\mathbf{s}}(c)$, then it can belong to only one Voronoi cell, viz. the nonstrict cell $\mathsf{Vo}(c)$ with the same nucleus, so that the right-hand side in (10.2.4) is a singleton set. Hence $\mathsf{Dig}(\{x\})$ is either empty or equal to the singleton set $\{p\}$. But if x belongs to, say, two Voronoi cells, the right-hand side in (10.2.4) consists of a set $\{p,q\}$ with $p \neq q$, and there is a choice: $\mathsf{Dig}(\{x\})$ may be equal to $\varnothing$, $\{p\}$, $\{q\}$, or $\{p,q\}$. And if x belongs to m Voronoi cells, the value can be any of 2^m subsets of Z.

Thus $\mathsf{Dig}(\{x\})$ is either empty or a singleton set whenever x belongs to the union of all strict Voronoi cells, but in the complement of that union, the value of the function may be a set with several elements. In some situations we do make a choice and define $\mathsf{Dig}(\{x\})$ to be a singleton set by introducing a new criterion. In fact, we have already done so when we defined the Khalimsky topology. If $X = \mathbf{R}$ and $Z = \mathbf{Z}$, then the Voronoi cells are the intervals $\left[n - \frac{1}{2}, n + \frac{1}{2}\right]$ and the strict cells are the open intervals $\left]n - \frac{1}{2}, n + \frac{1}{2}\right[$, $n \in \mathbf{Z}$. It is clear that the digitization of a real number which is not of the form $n + \frac{1}{2}$ is the empty set or $\left\{\left\lfloor x + \frac{1}{2}\right\rfloor\right\}$. When $x = n + \frac{1}{2}$, we may choose $F(\{x\})$ to be $\varnothing$, $\{n\}$, $\{n+1\}$, or $\{n, n+1\}$. When we defined the Khalimsky topology, we chose $\{n\}$ for n even and $\{n+1\}$ for n odd. But this is of course only one of many admissible choices.

Example 10.2.2. We get examples of Voronoi digitizations by taking $C(p) = \mathsf{Vo}(p)$ or $C(p) = \mathsf{Vo}_{\mathbf{s}}(p)$. Sometimes it is possible to choose a cell in between these two so that the space is covered exactly once by the different cells; an example was already mentioned: if $X = \mathbf{R}^n$ and $Z = \mathbf{Z}^n$ we may choose $C(p) = \prod \left]p_j - \frac{1}{2}, p_j + \frac{1}{2}\right]$.

Example 10.2.3. The digitization used by Rosenfeld (1974) is a Voronoi digitization, since the cell $C(p)$ defined in (10.1.3) is contained in the Voronoi cell, which is $\mathsf{Vo}(p) = \{x \in \mathbf{R}^2; \|x-p\|_\infty \leqslant \frac{1}{2}\}$.

10.3. Digital lines

In $\mathbf{R}^2$ we know what a straight line is: it is a set of the form $\{(1-t)a+tb; t \in \mathbf{R}\}$, where a and b are two distinct points in the plane. And a *straight line segment* is a connected subset of that line. We shall consider closed segments of finite length only, and may then write them as $\{(1-t)a+tb; 0 \leqslant t \leqslant 1\}$, where a and b are the endpoints. We shall denote this segment by $[a,b]$.

We shall choose $Z = \mathbf{Z}^2$ in the discussion that follows. The digitization of a straight line segment is the image under Dig of $[a,b]$, thus

$$\mathsf{Dig}([a,b]) = \bigcup_{t\in[0,1]} F(\{(1-t)a+tb\}) \subset \mathbf{Z}^2.$$

Suppose we are dealing with a Voronoi digitization. When $x = (1-t)a+tb$ belongs to a strict Voronoi cell, which in this case is $\mathsf{Vo}_\mathsf{s}(p) = \{x; \|x-p\|_\infty < \frac{1}{2}\}$, $p \in \mathbf{Z}^2$, then

$$F\big(\{x\}\big) = \big\{\big(\lfloor x_1 + \tfrac{1}{2}\rfloor, \lfloor x_1 + \tfrac{1}{2}\rfloor\big)\big\},$$

the unique point in $\mathbf{Z}^2$ closest to x. However, when x_1 is a half-integer, and x_2 is not, the digitization may be empty or consist of one or two points; when both coordinates are half-integers, the value may be a set of zero, one, two, three or four points.

In his famous paper (1974), Azriel Rosenfeld defined the digitization as in (10.1.3). In particular a point is always mapped to a point. For straight lines with slope less than 45°, he considered the intersections of its line segments with the vertical grid lines only. However, a line segment may intersect a horizontal grid line but no vertical grid line at all. In this case the cell is just the first segment in the union (10.1.3), but it does not matter so much, since the result will be trivially true for empty digitizations and the digitization is nonempty anyway for sufficiently long line segments.

We shall say with Rosenfeld that a subset A of $\mathbf{R}^2$ has *the chord property* if for all points $a, b \in A$ the segment $[a,b]$ is contained in $A + B_<(0,1)$, the dilation of A by the open unit ball (or disk or square) for the l^∞ metric.

Theorems 10.3.1 and 10.3.4 below are due to Rosenfeld (1974) and give together a characterization of the digitization of a straight line segment. The proof of Theorem 10.3.4 is new and is much shorter than the original proof.

Theorem 10.3.1. *The Rosenfeld digitization of a straight line segment has the chord property.*

Example 10.3.2. Let A be the set consisting of the five points $(0,0)$, $(1,0)$, $(2,0)$, $(3,1)$, $(4,2)$. This set does not have the chord property. Indeed, the point $(2,1)$ belongs to the segment $[(0,0),(4,2)]$, but it does not belong to the dilated set $A + B_<(0,1)$, although it does belong to the closed set $A + B_\leqslant(0,1)$. Thus, in view of the theorem, it cannot be the Rosenfeld digitization of a straight line segment. However, we may define a Voronoi digitization by declaring the digitization of $\big(0, -\frac{1}{2}\big)$ to be $(0,0)$, that of $\big(2, \frac{1}{2}\big)$ to be $(2,0)$, and that of $\big(4, 1\frac{1}{2}\big)$ to be $(4,2)$. Then A is the digitization of the straight line segment $\big[\big(0, -\frac{1}{2}\big), \big(4, 1\frac{1}{2}\big)\big]$. This digitization does not commute with translations, which offers a kind of explanation—of course it should not be allowed to move up by one half from $\big(0, -\frac{1}{2}\big)$ and $\big(4, 1\frac{1}{2}\big)$ and down by one half from $\big(2, \frac{1}{2}\big)$. Rosenfeld avoided this by always moving down in the case of half-integers.

Example 10.3.3. Slightly more generally we consider a set A consisting of five or six points $(0,0)$, $(1,0)$, $(2,0)$, and $a = (a_1, a_2)$, (a_1-1, a_2-1), (a_1-2, a_2-2), where $a_1 \geqslant 4$, $a_2 \geqslant 2$. (If $a = (4,2)$ we get the former example.) Then for no choice of a does this set have the chord property. Indeed, if $a_2 \geqslant \frac{1}{2}a_1$, then the point $(2, 2a_2/a_1)$, which is on the segment $[(0,0),(a_1,a_2)]$, does not belong to $A + B_<(0,1)$; if on the other hand $a_2 \leqslant \frac{1}{2}a_1$, then $(a_1-2, (a_1-2)a_2/a_1)$ on the same segment does not belong to $A + B_<(0,1)$.

Proof of Theorem 10.3.1. If the digitization of a line L has the chord property, so does the digitization of every segment of L. We may therefore restrict attention to the case of a whole line L. Let L be a straight line and $D \subset \mathbf{Z}^2$ its digitization. Let p, q be two points in D, and r an arbitrary point on the segment $[p,q]$. We shall prove that there exists a point $d \in D$ such that $\|d-r\|_\infty < 1$.

First we reduce to the case when the slope of L is between 0 and 1—note that the hypothesis and the conclusion are invariant under reflection and permutation of the coordinates. When it is exactly 0 or 1 the result is easy.

Instead of the digitization defined by (10.1.3) we shall now use only the vertical part of the cell,

$$C_{R,v}(p) = \big\{x; x_1 = p_1 \text{ and } p_2 - \tfrac{1}{2} < x_2 \leqslant p_2 + \tfrac{1}{2}\big\}. \tag{10.3.1}$$

When the slope of a line is strictly between 0 and 1, C_R and $C_{R,v}$ yield the same result.

When the slope of L is strictly between 0 and 1 we consider first the case when $r_1 \in \mathbf{Z}$. In this case we define $s \in \mathbf{R}^2$ as the point in L with $s_1 = r_1$. The digitization $d = \mathsf{Dig}(\{s\}) \in D$ of s satisfies $\|d-s\|_\infty \leqslant \frac{1}{2}$, thus $\|d-r\|_\infty \leqslant \|d-s\|_\infty + \|s-r\|_\infty \leqslant 1$. But can equality occur here? No. If we analyze the definition of the digitization we find that

$$r_2 - \tfrac{1}{2} < s_2 \leqslant r_2 + \tfrac{1}{2} \tag{10.3.2}$$

because of corresponding inequalities for p and q with respect to points on L, and that $d_2 - \frac{1}{2} < s_2 \leqslant d_2 + \frac{1}{2}$. Combining the two inequalities we see that

$$d_2 - 1 < s_2 - \tfrac{1}{2} \leqslant r_2 < s_2 + \tfrac{1}{2} \leqslant d_2 + 1, \tag{10.3.3}$$

so that actually $|d_2 - r_2| < 1$, while $|d_1 - r_1| = 0$, thus $r \in B_<(d,1)$.

Next we consider the case $r_1 \notin \mathbf{Z}$; $m < r_1 < m+1$ for some integer m. We now define s, s' and s'' as the points on L such that $s_1 = r_1$, $s'_1 = m$ and $s''_1 = m+1$, and let d' and d'' be the digitizations of s' and s''. Concerning d' and d'' we must have

$d'_2 \leqslant d''_2 \leqslant d'_2 + 1$. We shall therefore look separately at the two cases $d''_2 = d'_2$ and $d''_2 = d'_2 + 1$.

In case $d''_2 = d'_2$ we have $d'_2 - \frac{1}{2} < s'_2, s''_2 \leqslant d'_2 + \frac{1}{2}$, so the same inequality follows also for s_2, since s_2 is between s'_2 and s''_2. Combining with (10.3.2) we see that

$$d'_2 - 1 < s_2 - \tfrac{1}{2} \leqslant r_2 < s_2 + \tfrac{1}{2} \leqslant d'_2 + 1, \tag{10.3.4}$$

and conclude that $r \in B_<(d', 1) \cap B_<(d'', 1)$.

In case $d''_2 = d'_2 + 1$ we must have $r_2 > d'_2 - 1$ and $r_2 < d''_2 + 1$ so that $r \in B_<(d', 1) \cup B_<(d'', 1)$. The theorem is now completely proved.

To prove a converse we shall need the concept of digital arc. Let us say that two points in $\mathbf{Z}^2$ are *eight-neighbors* if their l^∞ distance is 1. Then a *digital arc* is a mapping from a finite integer interval $[a, b]_\mathbf{Z}$ into the plane $\mathbf{Z}^2$ which is Lipschitz-1 for the l^∞-norm and such that $\gamma(a)$ and $\gamma(b)$ have one eight-neighbor and $\gamma(x)$ has two eight-neighbors for $x = a + 1, \dots, b - 1$.

Theorem 10.3.4. *If a digital arc D in $\mathbf{Z}^2$ has the chord property, then it is the Rosenfeld digitization of some straight line segment in $\mathbf{R}^2$.*

Lemma 10.3.5. *Denote by $\pi_j \colon \mathbf{Z}^2 \to \mathbf{Z}$ the projection $(x_1, x_2) \mapsto x_j$, $j = 1, 2$. If a digital arc D has the chord property, then one of the restrictions $\pi_j\big|_D \colon D \to \mathbf{Z}$, $j = 1, 2$, is injective.*

Proof. Since D is a finite set it is contained in a minimal rectangle $[p_1, q_1] \times [p_2, q_2]$. If $p_1 = q_1$ or $p_2 = q_2$ we are done, so assume that $p_1 < q_1$ and $p_2 < q_2$. We claim that $\pi_1\big|_D$ is injective if $q_1 - p_1 \geqslant p_2 - q_2$; otherwise $\pi_2\big|_D$ is injective. So assume that $q_1 - p_1 \geqslant q_2 - p_2 > 0$. Each side of the rectangle must contain an endpoint of the arc; otherwise it cannot have the chord property. Since there are only two endpoints, they must be mapped to the vertices of the rectangle. After a possible reflection of the coordinates we may assume that the endpoints are $\gamma(a) = (p_1, p_2) = p$ and $\gamma(b) = (q_1, q_2) = q$. We claim that there are no two points on the arc with the same abscissa. If this were so, there would exist two such points with distance 1: $s = (s_1, s_2)$ and $t = (t_1, t_2)$ with $t_1 = s_1$ and $t_2 = s_2 + 1$. The point t cannot be an endpoint—that would violate the chord property for the segment $[p, t]$ and the point $r \in [p, t]$ with $r_1 = t_1 - 1$. Therefore t has a second neighbor in addition to s. But then this other neighbor must be $t' = (t_1 + 1, t_2 + 1)$, which violates the chord property for the segment $[p, t']$ and the point $r' \in [p, t']$ with $r'_1 = t_1 - 1$. This contradiction proves the lemma.

Proof of Theorem 10.3.4. Let D be a digital arc with the chord property. In view of the lemma and the symmetry of the digitization procedure, we may assume that there are no pairs of points a, b in D with $a_1 = b_1$, $a_2 \neq b_2$. Given three real numbers α, β, γ we define a strip in the plane by

$$S(\alpha, \beta, \gamma) = \{x \in \mathbf{R}^2; \alpha x_1 + \beta \leqslant x_2 \leqslant \alpha x_1 + \gamma\}.$$

Let us define the height of the strip as $\gamma - \beta$. The boundary $\partial S(\alpha, \beta, \gamma)$ of the strip has two components, given by the straight lines $x_2 = \alpha x_1 + \beta$ and $x_2 = \alpha x_1 + \gamma$. A finite set D of integer points is a subset of the digitization of a non-vertical straight line segment if and only if D is contained in a strip of height strictly less than 1.

For every given α there is a smallest strip $S(\alpha, \beta, \gamma)$ containing D. Moreover, varying also α, there is a strip $S_0 = S(\alpha_0, \beta_0, \gamma_0)$ of smallest height. If D consists of only one or two points, the conclusion follows easily, so let us assume that D has at least three points. Clearly there must be at least one point of D in each component of the boundary of S_0; otherwise we could increase β or decrease γ to obtain a narrower strip. And one of these lines must contain a second point of D; otherwise we could rotate the line slightly to obtain a strip of smaller height. For definiteness we shall assume that the three points on the boundary of the strip are p, s, q with $p_1 < s_1 < q_1$ and where p and q are on the lower boundary and s on the upper boundary. Let r be the point on $[p, q]$ with abscissa equal to that of s. (We note that p, s, q belong to $\mathbf{Z}^2$, while r need not do so.)

Now assume that D is not a subset of the digitization of a straight line. Then the height of this smallest strip is at least 1, so that $s_2 \geqslant r_2 + 1$, showing that r does not belong to $B_<(s, 1)$. To see that D does not satisfy the chord property we must however show that there is no $d \in D$ such that $r \in B_<(d, 1)$. So far we only know that r does not belong to $B_<(s, 1)$. However, s is the only point in D on the vertical line $x_1 = s_1$ and all other points $d \in D$ satisfy $|d_1 - r_1| = |d_1 - s_1| \geqslant 1$, so that $\|r - d\|_\infty \geqslant |r_1 - d_1| \geqslant 1$. Therefore D does not satisfy the chord property.

We have thus proved that a digital arc D having the chord property is a subset of the digitization of some straight line L. However, since D is a digital arc, it is the digitization of a connected subset of L. Obviously this subset can be taken to be compact, i.e., a straight line segment.

Melin (2003: Theorem 5) has proved a modification of this result when $\mathbf{Z}^2$ is given the Khalimsky topology and the digitization of any real line or line segment is homeomorphic to the Khalimsky line or a Khalimsky interval, respectively. [In a later version of these notes I would like to include his result.]

Generalize to other Voronoi digitizations....

Generalize to digitizations of convex sets....

10.4. Exercises

10.1. Let $X = \mathbf{R}^2$, $Z = \{(0,0), (a_1, a_2)\}$ and determine the Voronoi cells,
(a) when the metric is the Euclidean metric l^2 determined by the norm $\|\cdot\|_2$;
(b) when the metric is the l^∞ metric;
(c) when the metric is the l^1 metric.

10.2. Determine the Voronoi cells when $Z \subset \mathbf{C}$ is the set of all complex numbers $m + n\omega$, $m, n \in \mathbf{Z}$, where $\omega = \frac{1}{2} + \frac{i}{2}\sqrt{3}$ and we use the l^2 metric. What about other metrics?

References

Alexandroff, Paul (Aleksandrov, P. S.)

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

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

Baclawski, Kenneth; Björner, Anders

1979 Fixed points in partially ordered sets. *Advances in Mathematics* **31**, 263—287.

Birkhoff, Garrett

1940 *Lattice Theory.* New York City: American Mathematical Society. (Revised Edition 1948.)

Borgefors, Gunilla

1984 Distance transformations in arbitrary dimensions. *Comput. Vision Graphics Image Process.* **27**, 321—345.

1986 Distance transformations in digital images. *Comput. Vision Graphics Image Process.* **34**, 344—371.

1994 Applications using distance transforms. In: *Aspects of Visual Form Processing*, pp. 83—108, cf. (C. Arcelli; S. di Baja, Eds.), World Scientific, Singapore.

1996 On digital distance transforms in three dimensions. *Computer Vision and Image Understanding* **64**, 368—376.

Bourbaki, Nicolas

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

1963 *Théorie des ensembles.* Éléments de mathématique, première partie, livre I, chapitre 3. Second edition. Paris: Hermann.

Das, P. P.; Chatterji, B. M.

1988 Knight's distance in digital geometry. *Pattern Recognition Letters* **7**, 215—226.

Everett, C. J.

1944 Closure operators and Galois theory in lattices. *Trans. Amer. Math. Soc.* **55**, 514—525.

Gonzalez, Rafael C.; Woods, Richard E.

1993 *Digital Image Processing.* Addison-Wesley Publishing Company. xvi + 716 pp.

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. (Efim Khalimsky)

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

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

Herman, Gabor T.

1998 *Geometry of Digital Spaces.* Birkhäuser. x + 216 pp.

Hilditch, J.; Rutovitz, D.

1969 Chromosome recognition. *Annals of the New York Academy of Sciences* **157**, 339—364.

Hiriart-Urruty, Jean-Baptiste; Lemaréchal, Claude

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

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

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

Kiselman, Christer O.

1969 Prolongement des solutions d'une équation aux dérivées partielles à coefficients constants. *Bull. Soc. Math. France* **97**, 329—356.

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.

2003 La geometrio de la komputila ekrano. In: *Internacia Kongresa Universitato*, pp. 1—12. (Ed. Michela Lipari.) Rotterdam: Universala Esperanto Asocio. 83 pp.

ms Datorskärmens geometri. [The geometry of the computer screen. In Swedish.] Manuscript intended for the forthcoming book *Matematikens rikedomar.* 7 pp.

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.

Kopperman, R. D.; Kronheimer, E. H.; Wilson, R. G.

1998 Topologies on totally ordered sets. *Topol. Appl.* **90**, 165—185.

Kronheimer, E. H.

1992 The topology of digital images. *Topol. Appl.* **46**, 279—303.

Kuroš, A. G.

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

Marchand-Maillet, Stéphane; Sharaiha, Yazid M.

2000 *Binary Digital Image Processing.* Academic Press. xxv + 251 pp.

Matheron, G.

1967 *Éléments pour une théorie des milieux poreux.* Paris: Masson et C$^{\text{ie}}$.

1975 *Random sets and Integral Geometry.* New York: John Wiley & Sons. xxiii + 261 pp.

1988 Filters and lattices. *Image Analysis and Mathematical Morphology. Volume 2: Theoretical Advances*, 115—140. Jean Serra (Ed.). Academic Press.

Melin, Erik

2003 Digital straight lines in the Khalimsky plane. Uppsala: Uppsala University, Report 2003:30. To appear in *Mathematica Scandinavica.*

Montanari, U.

1968 A method for obtaining skeletons using a quasi-Euclidean distance. *J. Assoc. Comput. Mach.* **15**, 600—624.

Moore, Eliakim Hastings

1910 *Introduction to a Form of General Analysis.* New Have: Yale University Press.

Moreau, J.-J.

1970 Inf-convolution, sous-additivité, convexité des fonctions numériques, *J. Math. Pures et Appl.* **49**, 109—154.

Ore, Oystein

1944 Galois connexions. *Trans. Amer. Math. Soc.* **55**, 493—513.

Pfaltz, J. L.; Rosenfeld, A.

1967 Computer representation of planar regions by their skeletons. *Comm. ACM* **10**, 119—125.

Ragnemalm, I.,

1993 The Euclidean distance transform in arbitrary dimensions. *Pattern Recognition Letters* **14**, 883—888.

Rockafellar, R. Tyrrell

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

Rosenfeld, Azriel

1974 Digital straight line segments. *IEEE Transactions on Computers*, **c-23**, No. 12, 1264—1269.

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

Rosenfeld, A.; Pfaltz, J. L.

1966 Sequential operations in digital picture processing. *Journal of the ACM* **13**, 471-–494.

1968 Distance functions in digital pictures. *Pattern Recognit.* **1**, 33—61.

Serra, Jean

1982 *Image Analysis and Mathematical Morphology.* Academic Press. xviii + 610 pp.

1988 Mathematical morphology for complete lattices. *Image Analysis and Mathematical Morphology. Volume 2: Theoretical Advances*, 13—35. Jean Serra (Ed.). Academic Press.

1998 Connectivity on complete lattices. *Journal of Mathematical Imaging and Vision.* **9** (3), 231—251.

2001 *Lecture Notes on Morphological Operators.* Institut Mittag-Leffler, Lecture Notes No. 1, 2000/2001.

Singer, Ivan

1997 *Abstract Convex Analysis.* New York: John Wiley and Sons, Inc. xxii + 491 pp.

Starovoitov, Valery

1995 Toward a distance transform generalization. In: *Proceedings of the 9th Scandinavian Conference on Image Analysis*, pp. 499—506, (G. Borgefors, Ed.). Uppsala.

Strömberg, Thomas

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

Tarski, Alfred

1955 A lattice-theoretical fixpoint theorem and its applications. *Pacific J. Math.* **5**, 285—309.

Thiel, Edouard

2001 *Géométrie des distances de chanfrein.* Mémoire scientifique; Habilitation à diriger des recherches. Marseille: Université de la Méditerranée (Aix-Marseille II). 100 pp.

Tucker, A. W.

1936 Cell spaces. *Ann. of Math.* **37**, 92—100.

Verwer, Ben J. H.

1991 Local distances for distance transformations in two and three dimensions. *Pattern Recognition Letters*, **12**, 671—682.

Wyse, Frank, et al.

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

Index of terms

Index of symbols

Author's address:
Uppsala University, P. O. Box 480,
SE-751 06 Uppsala, Sweden.
E-mail:
`kiselman@math.uu.se`
URL:
`http://www.math.uu.se/~kiselman`
Telephone:
+46—18-4713216 (office);
+46—18-300708 (home);
+46—708-870708 (cellular).
Fax:
+46—18-4713201 (office).