Digital Geometry and Mathematical Morphology
Part I: Digital Geometry

Christer Kiselman

Svenska sällskapet för automatiserad bildanalys
Swedish Society for Automated Image Analysis
SSBA
Summer School in Uppsala
2007-08-14—17

Contents

Discrete models vs. models based on real numbers. Why digital geometry?

Distance transforms. Infimal convolution. Chamfer distances. Comparing distances.

Skeletons.

[Inverses and quotients of mappings between complete lattices.]

Digitization. Digital straight lines and planes as Diophantine approximations of real straight lines and planes. Chord properties.

Convexity. Digital straight lines and planes as convex sets.

[Discrete optimization.]

Topology. The Khalimsky line. The Khalimsky plane. Khalimsky straight line segments and planes. Khalimsky curves and surfaces.

Discrete models vs. models based on real numbers

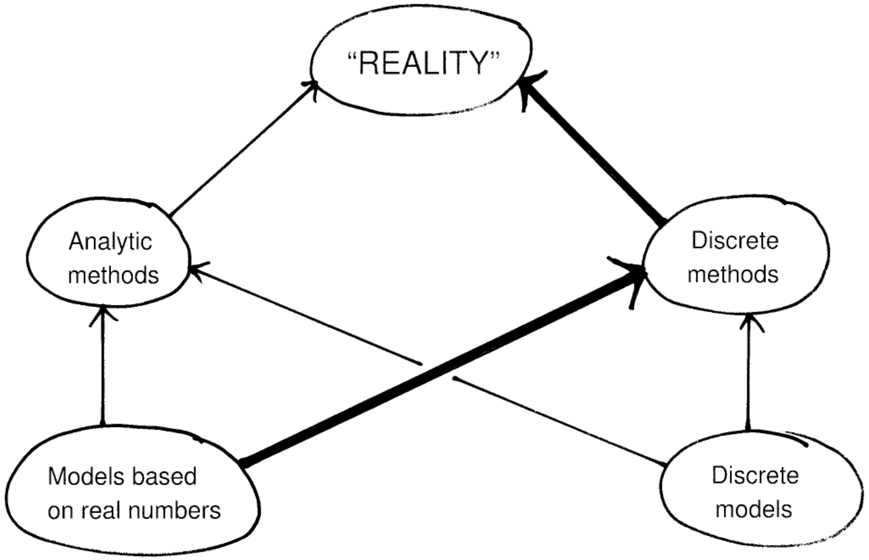
Mathematical models based on the real or complex numbers have been extremely successful in the sciences during several centuries (Newton and Leibniz).

How can the real numbers be a model of reality?

Physics with a smallest positive distance.

Real numbers versus discrete spaces.

Real numbers versus *p*-adic numbers.



$$f(x)=x^5,\quad x\in\mathbb{R},\qquad f'(x)=5x^4$$

$$\frac{f(x+h)-f(x)}{h}=5x^4+10x^3h+10x^2h^2+5xh^3+h^4$$

$$f(x+1)-f(x)=5x^4+10x^3+10x^2+5x+1$$

$$\int_0^x t^3dt=\frac{x^4}{4};$$

$$\sum_{t=0}^{x}t^3=\frac{x^4}{4}+Ax^3+Bx^2+Cx+D=\frac{x^4}{4}+\frac{x^3}{2}+\frac{x^2}{4}.$$

This formula was shown on South Korean Television, EBS, channel 48, 2006-03-15 06:30. And Japan, Channel 9, 2007-08!

$$\sum_{t=0}^{x}t^7=\frac{1}{24}x^2(x+1)^2(3x^4+6x^3-x^2-4x+2)$$

$$\sum_{t=0}^{x}t^m=\frac{x^{m+1}}{m+1}+\frac{x^m}{2}+\sum_{k=2}^{m}\binom{m}{k-1}\frac{B_k}{k}x^{m-k+1}.$$

Bernoulli numbers: $B_1=1/2$, $B_2=1/6$, $B_4=-1/30$, $B_6=1/42$; $B_3=B_5=B_7=\cdots=0$ (Gradštejn & Ryžyk, p. 15, 16).

Compare with

$$\int_0^x t^mdt=\frac{x^{m+1}}{m+1}.$$

A final example:

$$(fg)'=f'g+fg',\qquad (f*g)'=f'*g=f*g'$$

for functions defined on the real line. The second holds without change for functions on the integer line $\mathbb{Z}$, while the first is very problematic to translate to $\mathbb{Z}$. (Drell, Weinstein & Yankielowicz 1976, Bouguenaya & Fairlie 1986).

All this can be read in two different ways:
(1) as propaganda for differential and integral calculus;
(2) as a challenge.

Thinking about this Summer School on board the FinnEagle 2007-07-13: Has the whole world gone digital?

Why digital geometry?

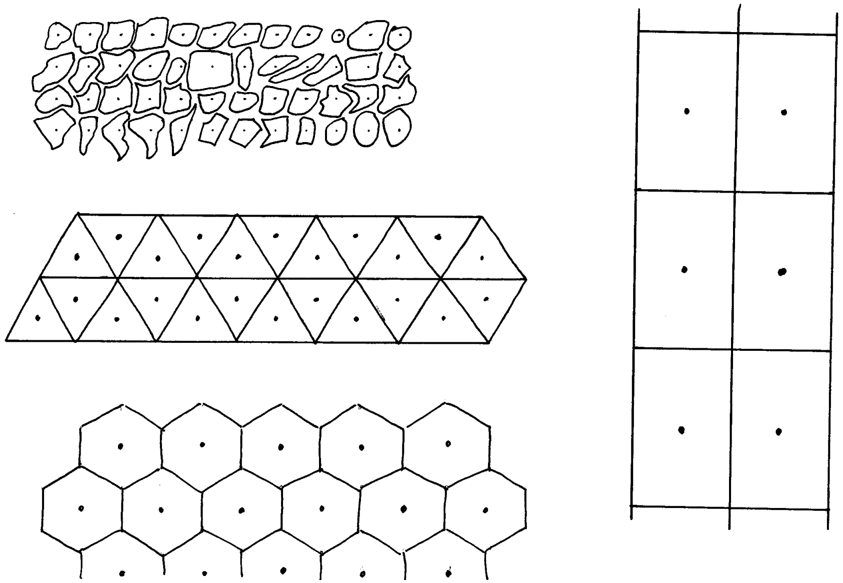
The geometry of the computer screen.

Points, straight line segments, planes. Ellipses, hyperbolas. Lemniscates, cardioids.

Euclid: ευθεῖα, *eutheĩa* ‘straight line segment, rectilinear segment’. Between two points on a line there is always a third point (hence infinitely many). There are no equilateral triangles in $\mathbb{Q}^2$.

Digital objects might be viewed as approximations of Euclidean objects. But it is better to treat them for what they are! Finite sets of objects!

Rosenfeld 1974. Tessellations of the Euclidean plane. Pixels, voxels! Adresses of pixels, voxels: $\mathbb{Z}^2$, $\mathbb{Z}^3$.



Distance transforms

Distance transforms of digital images: 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 we shall define it as minus the distance to the complement. The choice of signs has a very simple motivation: the distance transform of a Euclidean ball $B = \{x \in \mathbb{R}^2; \|x-c\|_2 \leqslant r\}$ is $\mathbf{DT}_B(x) = \|x-c\|_2 - r$, a convex function. More generally, the distance transform of a convex set in $\mathbb{R}^n$ 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: 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 $\mathbb{Z}^n$ later.

Let X be any nonempty set. A function $d\colon X \times X \to \mathbb{R}$ is called a ***distance*** if d is ***positive definite***, i.e.,

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

and ***symmetric***, i.e.,

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

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

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

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

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

Symmetry? A cost function need not be symmetric...

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

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

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

A ***metric space*** is simply a set provided with a metric.

The ***strict ball*** (or ***open ball***) of center c and radius r

$$B_<(c,r) = \{x; d(c,x) < r\}.$$

The ***non-strict ball*** (or ***closed ball***) of center c and radius r is

$$B_\leqslant(c,r) = \{x; d(c,x) \leqslant r\} = B_<(c,r) \cup \{x; d(c,x) = r\}.$$

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 $B_<(c,r)$.

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

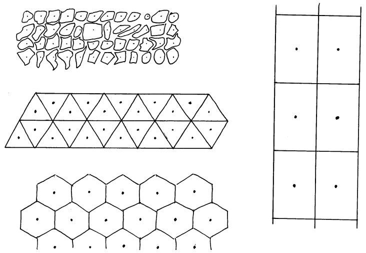
$$B_<(c_1,r_1) + B_<(c_2,r_2) \subset B_\leqslant(c_1,r_1) + B_<(c_2,r_2) \subset B_<(c_3,r_3);$$

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

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

Tessellations

The plane may be divided into triangles, rectangles, or hexagons. These are the most common tessellations of the plane. The centers of the pixels form, respectively, a hexagonal, rectangular, or triangular pattern. (M. C. Escher; Gunilla Borgefors.)



Definition

Let A and B be subsets of an abelian group G. Then we define their ***Minkowski sum*** as the set

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

If B is finite, as is often the case in $\mathbb{Z}^n$, 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 $b\in B$.

If B is a singleton set, $B=\{b\}$, we may write $A+\{b\}=A+b$.

Definition

Dilation by C, $\delta_C\colon \mathscr{P}(G)\to\mathscr{P}(G)$, and ***erosion*** by C, $\varepsilon_C\colon \mathscr{P}(G)\to\mathscr{P}(G)$, are defined by

$$\delta_C(A)=A+C, \qquad \varepsilon_C(A)=\{x; x+C\subset A\}.$$

For symmetry they may be written

$$\delta_C(A)=\bigcup_x\big(\{x\}+C; \{x\}\subset A\big), \qquad \varepsilon_C(A)=\bigcup_x\big(\{x\}; \{x\}+C\subset A\big).$$

There are two kinds of duality between the two operations:

Proposition

(Group-theoretical duality.) *Define $\check{C}=\{-c; c\in C\}$. Then*

$$\delta_C(A)=\complement\varepsilon_{\check{C}}\left(\complement A\right).$$

Proposition

(Lattice-theoretical duality.) *Let A, B and C be three subsets of an abelian group. Then $\delta_C(A)\subset B$ if and only if $A\subset\varepsilon_C(B)$.*

Dilation commutes with the formation of unions, and erosions with the formation of intersections:

$$\delta_C\left(\bigcup_{j\in J}A_j\right)=\bigcup_{j\in J}\delta(A_j), \qquad \varepsilon_C\left(\bigcap_{j\in J}A_j\right)=\bigcap_{j\in J}\varepsilon(A_j).$$

These properties are taken as definitions in the lattice-theoretical approach to dilations and erosions.

Given two functions $f,g\colon G\to[-\infty,+\infty]$ with values in the ***extended real line*** $[-\infty,+\infty]=\mathbb{R}\cup\{-\infty,+\infty\}$, we define a new function $h=f\mathbin{\square}g$, called the ***infimal convolution*** of f and g, as

$$(f\mathbin{\square}g)(z)=h(z)=\inf_{x,y\in G}\Big(f(x)\dotplus g(y); x+y=z\Big), \qquad z\in G.$$

The infimum is taken over all elements $x,y\in G$ such that their sum is z, the argument of h.

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 $s\dotplus t$ 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 $[-\infty,+\infty]$.

Similarly we define ***lower addition***, $s\underset{\cdot}{+}t=-\big((-s)\dotplus(-t)\big)$; here minus infinity wins.

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:

$$(f \mathbin{\square} g)(z) = \inf_{x,y\in G}(f(x)+g(y); x+y=z, f(x)<+\infty, g(y)<+\infty).$$

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:

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

With this notation we can write

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

Intuitively, plus infinity corresponds to vacuum and $-\infty$ to an infinitely dense neutron star. We should think of the density as $e^{-f(x)}$, and then of course $e^{-(+\infty)} = 0$. Infimal convolution is related to 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 $\mathbb{R}^n$, which means that we sometimes have a remarkably good approximation

$$e^{(f \mathbin{\square} g)(x)} = \sup_{y\in\mathbb{R}^n} \left[e^{-f(y)}e^{-g(x-y)}\right]$$

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

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

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

Example

Define for $a > 0$,

$$f_a(x) = \frac{x^2}{2a}, \qquad x \in \mathbb{R}.$$

Then

$$f_a \mathbin{\square} f_b = f_{a+b}$$

and

$$e^{-f_a} * e^{-f_b} = C_{a,b}e^{-f_{a+b}},$$

where

$$C_{a,b} = \sqrt{\frac{2\pi ab}{a+b}}.$$

Proposition

Infimal convolution is associative: $(f_1 \mathbin{\square} f_2) \mathbin{\square} f_3 = f_1 \mathbin{\square} (f_2 \mathbin{\square} f_3)$.

Now why is infimal convolution more general than Minkowski addition? This is because of the formula

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

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***, $\mathbf{ind}_A$, which is simply defined as $\mathbf{ind}_A(x) = 0$ when $x \in A$ and $\mathbf{ind}_A(x) = +\infty$ when $x \notin A$. It is related to the ***characteristic function*** χ_A of A by the formula $\chi_A = \exp(-\mathbf{ind}_A)$.

It is clear that $\operatorname{dom}(\mathbf{ind}_A) = A$. We have $\mathbf{ind}_A \square \mathbf{ind}_B = \mathbf{ind}_{A+B}$ for all subsets A, B of an abelian group G. Hence the Minkowski sum may be defined in terms of infimal convolution as $A+B = \operatorname{dom}(\mathbf{ind}_A \square \mathbf{ind}_B)$.

We can go also in the other direction.

The ***epigraph*** of a function $f\colon X \to [-\infty, +\infty]$ defined on an arbitrary set X is

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

and the ***strict epigraph*** is

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

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

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

This means that the function $f \square g$ can be defined as the function whose strict epigraph is the sum $(\operatorname{epi_s} f) + (\operatorname{epi_s} g)$.

Distance transforms

Definition

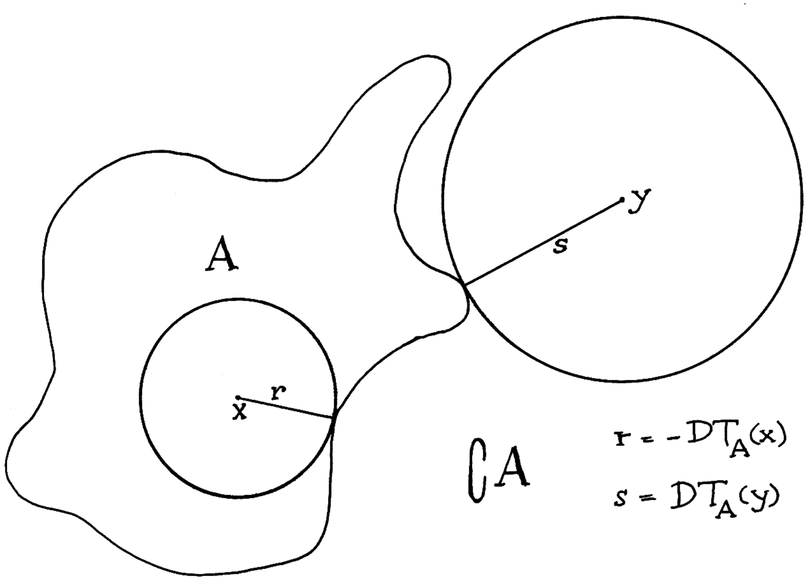
In a metric space X we define the ***distance transform*** $\mathbf{DT}_A$ of a subset A of X by

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

Lemma

The distance transform satisfies

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



Note the symmetry: $\mathbf{DT}_{X \smallsetminus A} = -\mathbf{DT}_A$. The distance transformation $A \mapsto \mathbf{DT}_A$ is decreasing in the sense that $\mathbf{DT}_A(x) \geqslant \mathbf{DT}_B(x)$ for all $x \in X$ if $A \subset B$. In the two extreme cases $A = \varnothing$ and $A = X$ we have $\mathbf{DT}_\varnothing = +\infty$ and $\mathbf{DT}_X = -\infty$. In all other cases $\mathbf{DT}_A$ is real-valued, $\mathbf{DT}_A\colon X \to \mathbb{R}$.

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, $\mathbf{DT}_A = (\mathbf{DT}_A)^+ - (\mathbf{DT}_A)^-$. The function $(\mathbf{DT}_A)^-$ is sometimes called the ***quench function*** of A.

Proposition

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

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

and similarly for $(\mathbf{DT}_A)^-$. (In particular the restrictions $\mathbf{DT}_A\big|_A$ and $\mathbf{DT}_A\big|_{\complement A}$ are Lipschitz continuous with Lipschitz constant 1.) As a consequence, $\mathbf{DT}_A$ 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.

Proof. The restriction to A of $\mathbf{DT}_A$ is the supremum of a family of Lip-1 functions $x\mapsto d(x,y)$. Hence $|(\mathbf{DT}_A)^-(a)-(\mathbf{DT}_A)^-(b)|\leqslant d(a,b)$ 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 $r=-\mathbf{DT}_A(a)\geqslant 0$ and $s=\mathbf{DT}_A(b)\geqslant 0$. Then $B_<(a,r)\subset A$ and $B_<(b,s)\subset\complement A$, 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 $0\leqslant \mathbf{DT}_A(b)-\mathbf{DT}_A(a)=s+r\leqslant d(a,b)$; proving that the Lipschitz constant is 1 in this case.

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

The Lipschitz constant 2 in the proposition cannot be improved: take $X=\mathbb{Z}$ 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 $X\smallsetminus A$. 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.

Proposition

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. Then

$$(\mathbf{DT}_A)^+=\max(\mathbf{DT}_A,0)=\mathbf{ind}_A\mathbin{\square} f$$

and

$$(\mathbf{DT}_A)^-=\max(-\mathbf{DT}_A,0)=\mathbf{ind}_{\complement A}\mathbin{\square} f,$$

and, taking the difference between the two,

$$\mathbf{DT}_A=(\mathbf{DT}_A)^+-(\mathbf{DT}_A)^-=(\mathbf{ind}_A\mathbin{\square} f)-(\mathbf{ind}_{\complement A}\mathbin{\square} f).$$

Proposition

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

$$(\mathbf{DT}_{A+B})^+=(\mathbf{DT}_A)^+\mathbin{\square}\mathbf{ind}_B=\mathbf{ind}_A\mathbin{\square}(\mathbf{DT}_B)^+=(\mathbf{DT}_A)^+\mathbin{\square}(\mathbf{DT}_B)^+.$$

Proof. We know from the preceding proposition that $(\mathbf{DT}_A)^+=\mathbf{ind}_A\mathbin{\square} f$, where $f(x)=d(x,0)$ is the distance from x to the origin. Hence, using the associativity and commutativity of infimal convolution as well as the functional equation $f\mathbin{\square} f=f$ (comes up soon),

$$\begin{aligned}(\mathbf{DT}_A)^+\mathbin{\square}(\mathbf{DT}_B)^+&=(\mathbf{ind}_A\mathbin{\square} f)\mathbin{\square}(\mathbf{ind}_B\mathbin{\square} f)=(\mathbf{ind}_A\mathbin{\square}\mathbf{ind}_B)\mathbin{\square} f\\&=\mathbf{ind}_{A+B}\mathbin{\square} f=(\mathbf{DT}_{A+B})^+.\end{aligned}$$

Also

$$\begin{aligned}(\mathbf{DT}_A)^+\mathbin{\square}\mathbf{ind}_B&=(\mathbf{ind}_A\mathbin{\square} f)\mathbin{\square}\mathbf{ind}_B=(\mathbf{ind}_A\mathbin{\square}\mathbf{ind}_B)\mathbin{\square} f\\&=\mathbf{ind}_{A+B}\mathbin{\square} f=(\mathbf{DT}_{A+B})^+.\end{aligned}$$

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 $[-\infty,+\infty]$. We shall denote them by $\{f < s\}$ rather than $\{x \in X; f(x) < s\}$.

Lemma

If X is a metric space with metric d, and $\mathbf{DT}_A$ 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 $\mathbf{DT}_A$:

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

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

If A is any subset of $\mathbb{R}^n$ 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 $\mathbb{R}^n$, then $\mathbf{DT}_A(x) = \|x-c\| - r$. This simple example shows that we cannot expect to recover A exactly from $\mathbf{DT}_A$; we have to be content with its interior and closure. However, if $X = \mathbb{Z}^n$, then the topology induced by a norm in $\mathbb{R}^n$ is the discrete topology, so that, for any set A,

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

The boundary is empty and $\mathbf{DT}_A$ never takes the value zero. □

Proposition

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

$$\{\mathbf{DT}_A < r\} = A + B_<(0,r) = \delta_{B_<(0,r)}(A) \subset A + B_\leqslant(0,r)$$

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

and

$$\{\mathbf{DT}_A \leqslant -r\} = \varepsilon_{B_<(0,r)}(A) \supset \varepsilon_{B_\leqslant(0,r)}(A) \supset \{\mathbf{DT}_A < -r\}$$

$$\supset \{\mathbf{DT}_A \leqslant -r - \varepsilon\}.$$

The dilations by the strict balls $B_<(0,r)$, $A + B_<(0,r) = \delta_{B_<(0,r)}(A)$, thus determine the strict sublevel sets of $\mathbf{DT}_A$ for positive values; similarly for the erosions $\varepsilon_{B_<(0,r)}(A)$ and the nonstrict sublevel sets of $\mathbf{DT}_A$ for negative values.

Proposition

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

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

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

Moreover, for any real number r_3 we have

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

When the functions f_j are distance transforms and r_3 is positive we can say more:

Proposition

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 $f_j = \mathbf{DT}_{A_j}$ satisfy

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

$$= A_1 + A_2 + B_<(0, r)$$

for all positive r.

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.

	1	
1	0	1
	1	

9	8	7	6	5	4	5	6	7	8	9
8	7	6	5	4	3	4	5	6	7	8
7	6	5	4	3	2	3	4	5	6	7
6	5	4	3	2	1	2	3	4	5	6
5	4	3	2	1	0	1	2	3	4	5
6	5	4	3	2	1	2	3	4	5	6
7	6	5	4	3	2	3	4	5	6	7
8	7	6	5	4	3	4	5	6	7	8
9	8	7	6	5	4	5	6	7	8	9

1	1	1
1	0	1
1	1	1

5	4	4	4	4	4	4	4	4	4	5
5	4	3	3	3	3	3	3	3	4	5
5	4	3	2	2	2	2	2	3	4	5
5	4	3	2	1	1	1	2	3	4	5
5	4	3	2	1	0	1	2	3	4	5
5	4	3	2	1	1	1	2	3	4	5
5	4	3	2	2	2	2	2	3	4	5
5	4	3	3	3	3	3	3	3	4	5
5	4	4	4	4	4	4	4	4	4	5

The following result is well known and easy to prove.

Lemma

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

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

and symmetric:

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

Conversely, a function f which satisfies these two conditions defines a distance $d(x, y) = f(x - y)$.

Note that we do not need the triangle inequality here. But that special case is easy to recognize:

Lemma

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

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

Often the infimum in an infimal convolution over $\mathbb{Z}^n$ 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 $\{y; g(y) \leqslant a, a \in \mathbb{R}\}$, 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 $g < +\infty$ in a finite set P only. Then the infimal convolution with any function f is equal to the minimum

$$(f \,\square\, g)(x) = \min_{y \in P} \big(f(x - y) \dot{+} g(y)\big), \qquad x \in G.$$

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

Lemma

A function f on an abelian group is subadditive if and only if it satisfies the inequality $f \,\square\, f \geqslant f$. If $f(0) = 0$, this is equivalent to the equation $f \,\square\, f = f$.

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

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

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 $-\infty$. It is natural to think of a path leading from 0 to x consisting of segments $[0, x^1]$, $[x^1, x^1 + x^2]$, $\dots$, $[x^1 + \cdots + x^{k-1}, x]$; if $G = \mathbb{Z}^2$ this path can be realized in $\mathbb{R}^2$.

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

$$\mathbb{N} \cdot A = \{\sum m_i a_i;\, m_i \in \mathbb{N}, a_i \in A\},$$

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

$$\mathbb{Z} \cdot A = \{\sum m_i a_i;\, m_i \in \mathbb{Z}, a_i \in A\}.$$

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

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

Theorem

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 $F_1 = F$, $F_j = F_{j-1} \mathbin{\square} F$, $j = 2, 3, \dots$, 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 $\operatorname{dom} f = \mathbb{N} \cdot \operatorname{dom} F$, i.e., f is finite precisely in the semigroup generated by $\operatorname{dom} F$.

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

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

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 $F_{j+k} = F_j \mathbin{\square} F_k$, so we get

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

Since ε is arbitrary, the inequality $f(x+y) \leqslant f(x) + f(y)$ follows. □

Theorem

With F as in the previous theorem, 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 $\mathbb{Z} \cdot P = \mathbb{N} \cdot P$ 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. We know that $H \,\square\, H = H$ 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. In $\mathbb{Z}^n$ we can also use $d_1(x,y) = \varepsilon\|x-y\|$ for any norm on $\mathbb{R}^n$.

Corollary

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 $\mathbb{Z} \cdot P = \mathbb{N} \cdot P$ 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 earlier. We can now apply the theorem. □

Definition

We shall say that a metric $d(x,y) = f(x-y)$ is a ***chamfer distance***, or ***finitely generated*** if it is constructed as in the corollary.

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

It is by no means necessary that f be positively homogeneous in the corollary.

1	7	2
8	0	8
2	7	1

12	4	10	5	11	6	12	7	13	8	16
6	11	3	9	4	10	5	11	6	14	9
13	5	10	2	8	3	9	4	12	7	15
7	12	4	9	1	7	2	10	5	13	8
14	6	11	3	8	0	8	3	11	6	14
8	13	5	10	2	7	1	9	4	12	7
15	7	12	4	9	3	8	2	10	5	13
9	14	6	11	5	10	4	9	3	11	6
16	8	13	7	12	6	11	5	10	4	12

Several metrics on $\mathbb{Z}^2$ have been studied. When presenting the generating function F defining the prime distances it shall be understood 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 $P = \{x \in \mathbb{Z}^2; \sum |x_j| \leqslant 1\}$ 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 $P = \{x \in \mathbb{Z}^2; |x_j| \leqslant 1\}$ 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 find explicitly an index j such that $F_j(x)$ is equal to the limit $f(x)$:

Proposition

Let F be as in the corollary. 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 $j \geqslant s(x)$.

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 $j = 1, \dots, 99$ but $f(100) = F_{100}(100) = 100$.

Corollary

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

What about $\lim(\mathbf{ind}_A \,\square\, F_1 \,\square\, F_2 \,\square\, F_3 \,\square\, \cdots \,\square\, F_k)$ with different F_j? Benedek Nagy and Robin Strand!

Comparing distances

The l^1 (city-block) and l^∞ (chessboard) metrics in $\mathbb{R}^2$ are translation invariant but not rotation invariant. (In the plane a rotation can distort distances by a factor of up to $\sqrt{2}$; in $\mathbb{R}^n$ by $\sqrt{n}$.) The Euclidean metric is rotation invariant, and it is desirable to construct a chamfer distance in $\mathbb{Z}^n$ 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. We shall look briefly into this problem and describe some methods of comparing two nonnegative functions.

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,

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

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 $\Lambda(f_1, g) \leqslant \sqrt{C_2/C_1}$.

Verwer (1991) used instead the functional

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

One might just as well consider $\Lambda'(g,f)$. Clearly $\Lambda'(f,g) = \Lambda(f,g) - 1$ when $f \geqslant g$, and $\Lambda'(f,g) = 1 - 1/\Lambda(f,g)$ 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 inserted also the well-known inequality $1 - 1/t \leqslant \log t \leqslant t - 1$. In particular, $|\Lambda'(f,g) - \log\Lambda(f,g)| \leqslant (\Lambda(f,g) - 1)^2/\Lambda(f,g)$.

We may also note that Λ' 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.

If the prime vectors are $(\pm 1, 0)$, $(0, \pm 1)$ and $(\pm 1, \pm 1)$ 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

$$a_0 = \sqrt[4]{\frac{2+\sqrt{2}}{4}} \approx 0.961186523, \quad b_0 = a_0\sqrt{2} = \sqrt[4]{2+\sqrt{2}} \approx 1.359323017.$$

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

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

The exact values are $a_1 = \left(\frac{1}{2} + \sqrt{1 - 1/\sqrt{2}}\right)^{-1}$ and $b_1 = a_1\sqrt{2}$.

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

$$a_2 = \tfrac{1}{2} + \frac{1}{4}\sqrt{2+\sqrt{2}} \approx 0.961939766, \quad b_2 = a_2\sqrt{2} \approx 1.3603882.$$

The vertices of the octagon protrude as much outside the disk as the midpoints of the edges go into the disk. As expected, $a_0 = \sqrt{a_1 a_2}$.

Skeletons

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 $r = -\mathbf{DT}_A(a)$. In fact, when we defined the distance transform $\mathbf{DT}_A(a)$ 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

Let A be a subset of a metric space X. We define the ***skeleton*** of A, denoted by $\mathbf{sk}(A)$, as the set of all centers of maximal nonempty strict balls contained in A.

The definition means that $a \in \mathbf{sk}(A)$ 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 $B_<(b, s) = B_<(a, r)$. The skeleton may be empty: think of a set with empty interior or of a half-space in $\mathbb{R}^n$. 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. To do this in $\mathbb{R}^n$, we shall need Zorn's Lemma.

Preorders and orders

Definition

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

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

$$x \leqslant x$$

and

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

Definition

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

This means that it satisfies: for all $x, y \in X$,

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

Definition

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 $y \leqslant x$.

Definition

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.

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

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

Theorem

Let $\mathbb{Z}^n$ be equipped with a metric which either is inherited from a norm on $\mathbb{R}^n$ or a chamfer distance, and let A be a finite subset. Then the set of all strict 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 $\mathbb{Z}^n \times \mathbb{R}$. Assume that the family is totally ordered, i.e., that 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 $\mathbb{R}^n$ things are less simple.

Theorem

Let A be a set in a finite-dimensional normed vector space E. Assume that A is bounded and has nonempty interior. Then the set of open balls contained in A is inductively ordered.

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

Corollary

Let A be a bounded subset of a finite-dimensional normed vector space, or a bounded subset of $\mathbb{Z}^n$, where $\mathbb{Z}^n$ is provided with a metric as in the theorem. The union of all open balls with center c belonging to the skeleton and radius equal to $-\mathbf{DT}_A(c)$ 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 and one of the two previous theorems, there is a maximal ball $B_<(c,r)$ containing $B_<(x,\varepsilon)$ and contained in A. Thus $c \in \mathbf{sk}(A)$ and $x \in B_<(c,r)$, with $r = -\mathbf{DT}_A(c)$. □

In any metric space where the conclusion of the theorem holds we have

$$A^\circ = \bigcup_{c \in \mathbf{sk}(A)} B_<(c, -\mathbf{DT}_A(c)).$$

Here $-\mathbf{DT}_A(c) = (\mathbf{DT}_A(c))^-$ is the quench function evaluated at c. Knowledge of $\mathbf{sk}(A)$ and the restricion of $\mathbf{DT}_A$ to $\mathbf{sk}(A)$ is equivalent to knowing A°. This shows how we can reconstruct A° from $\mathbf{sk}(A)$ 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.

In some sense the skeleton is a thin set. For instance, it is easy to prove that a skeleton in $\mathbb{R}^n$ has no interior points. On the other hand, the closure of the skeleton need not be of Lebesgue measure zero. These results are mentioned by Serra (1982:378) and Matheron (1988:218). Rivière (1996) proved that the skeleton is of Lebesgue measure zero. It is probably unknown whether the interior of its closure is empty.

The skeleton has, generally speaking, bad continuity properties.

Example

Let D be the open unit disk in $\mathbb{R}^2$, $D = \{(x,y) \in \mathbb{R}^2; x^2 + y^2 < 1\}$. 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 $[(0,0),(1,0)]$ 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 $(0,0)$ 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 $\mathbb{Z}^2$ the continuity properties are of course different, but a small change can still cause points to appear far from the original skeleton.

Example

Let $A = [-m, m]_{\mathbb{Z}} \times [-m, m]_{\mathbb{Z}}$ be a large square in $\mathbb{Z}^2$. 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 $A \smallsetminus \{(m, 0)\}$.

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

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

Theorem

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 $h(x) = d(x, c) + \mathbf{DT}_A(x)$, $x \in E$. Then c belongs to the skeleton of A if and only if h has a minimum only at c.

There are similar results for discrete spaces, but they are not so easy to describe.

Thanks to the calculus of balls we can generalize this result to other groups. In a normed space of positive dimension, the open ball of radius $r = -\mathbf{DT}_A(a)$ 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 $x \mapsto d(x, c) + \mathbf{DT}_A(x)$ by $x \mapsto d(x, c) - \rho_-(-\mathbf{DT}_A(x))$.

Theorem

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 $h(x) = d(x, c) - \rho_-(-\mathbf{DT}_A(x))$, $x \in G$. Then c belongs to the skeleton of A if and only if h has a minimum only at c.

Digitization

Let us now 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 $\mathbb{R}^2$ and Z as $\mathbb{Z}^2$ 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 the direct image of A under f, $f_*(A) = \{f(x); x \in A\}$.

However, 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 $Z \neq X$.

The mappings f_* are dilations in the lattice-theoretical sense: they commute with the formation of unions. 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

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

$$F(\{x\}) = \{p \in Z; x \in C(p)\}, \qquad x \in X,$$

and

$$F(A) = \bigcup_{x\in A} F(\{x\}) = \{p \in Z; A \text{ meets } C(p)\}, \qquad A \in \mathscr{P}(X).$$

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 $\mathscr{C}_A \in \mathscr{P}(\mathscr{P}(X))$ is the family of all cells which meet A. (In general, we define f^* for a mapping $f\colon X \to Y$ by $f^*(B) = \{x \in X; f(x) \in B\}$, the preimage of B.)

It is clear that a digitization in this sense is always a dilation in the lattice-theoretical sense. 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. We will see one soon.

Example

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

If $X = \mathbb{R}$ and $Z = \mathbb{Z}$ 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 $x \mapsto -x$.

If $X = \mathbb{R}^2$ and $Z = \mathbb{Z}^2$, we may construct digitizations from what we have already done on the real axis. We may take

$$C(p) = \left[p_1 - \tfrac{1}{2}, p_1 + \tfrac{1}{2}\right] \times \left[p_2 - \tfrac{1}{2}, p_2 + \tfrac{1}{2}\right], \qquad p \in \mathbb{Z}^2,$$

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

$$C_R(p) = \left\{x; x_1 = p_1 \text{ and } p_2 - \tfrac{1}{2} < x_2 \leqslant p_2 + \tfrac{1}{2}\right\}$$
$$\cup \left\{x; p_1 - \tfrac{1}{2} < x_1 \leqslant p_1 + \tfrac{1}{2} \text{ and } x_2 = p_2\right\}.$$

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 $\mathbb{R} \ni x \mapsto \left\lceil x - \frac{1}{2} \right\rceil \in \mathbb{Z}$ already mentioned, a digitization of $\mathbb{R}$ 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 $\mathbb{R}^2$.

It is clear that in this case the union of the cells is very small compared with $\mathbb{R}^2$, so that many sets have empty digitization. However, the union of all cells is equal to all grid lines $(\mathbb{R} \times \mathbb{Z}) \cup (\mathbb{Z} \times \mathbb{R})$, 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.

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*** with nucleus p is

$$\mathbf{Vo}(p) = \{x \in X; \forall q \in Z, d(x,p) \leqslant d(x,q)\}, \qquad p \in Z.$$

Thus $x \in \mathbf{Vo}(p)$ if and only if p is a best approximant of x. We also define the ***strict Voronoi cell*** as

$$\mathbf{Vo}_{\mathrm{s}}(p) = \{x \in X; \forall q \in Z \smallsetminus \{p\}, d(x,p) < d(x,q)\} \qquad p \in Z.$$

Thus $x \in \mathbf{Vo}_{\mathrm{s}}(p)$ if and only if p is the unique best approximant of x. Finally, one might define the ***very strict Voronoi cell*** as

$$\mathbf{Vo}_{\mathrm{vs}}(p) = \{x \in X; d(x,p) < \inf_{q \in Z \smallsetminus \{p\}} d(x,q)\}, \qquad p \in Z.$$

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, 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 consist only of 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

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 $r \in \mathbb{R}$. A ***Voronoi digitization of*** X ***into*** Z is a dilation $\mathbf{dig}\colon \mathscr{P}(X) \to \mathscr{P}(Z)$ such that

$$\mathbf{dig}(\{x\}) \subset \{p \in Z; x \in \mathbf{Vo}(p)\}.$$

Note that if x belongs to some strict Voronoi cell $\mathbf{Vo}_{\mathrm{s}}(c)$, then it can belong to only one Voronoi cell, viz. the nonstrict cell $\mathbf{Vo}(c)$ with the same nucleus, so that the right-hand side is a singleton set. Hence $\mathbf{dig}(\{x\})$ is either empty or equal to the singleton set $\{p\}$. But if x belongs to, say, two Voronoi cells, the right-hand side consists of a set $\{p,q\}$ with $p \neq q$, and there is a choice: $\mathbf{dig}(\{x\})$ may be equal to Ø, $\{p\}$, $\{q\}$, or $\{p,q\}$. And if x belongs to m Voronoi cells, the value can be any of 2^m subsets of Z.

Thus $\mathbf{dig}(\{x\})$ 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 $\mathbf{dig}(\{x\})$ to be a singleton set by introducing a new criterion. In fact, we shall do so when we define the Khalimsky topology. If $X = \mathbb{R}$ and $Z = \mathbb{Z}$, 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 $\left]n-\frac{1}{2}, n+\frac{1}{2}\right[$, $n \in \mathbb{Z}$.

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 Ø, $\{n\}$, $\{n+1\}$, or $\{n, n+1\}$. When we define the Khalimsky topology, we shall choose $\{n\}$ for n even and $\{n+1\}$ for n odd. But this is of course only one of many admissible choices.

Example

We get examples of Voronoi digitizations by taking $C(p) = \mathbf{Vo}(p)$ or $C(p) = \mathbf{Vo}_s(p)$. 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 $X = \mathbb{R}^n$ and $Z = \mathbb{Z}^n$ we may choose $C(p) = \prod \left]p_j - \frac{1}{2}, p_j + \frac{1}{2}\right]$.

Example

The digitization used by Rosenfeld (1974) is a Voronoi digitization, since the cell $C_R(p)$ is contained in the Voronoi cell, which is $\mathbf{Vo}(p) = \left\{x \in \mathbb{R}^2; \|x - p\|_\infty \leqslant \frac{1}{2}\right\}$.

Digital lines

We know what a straight line in $\mathbb{R}^2$ is: it is a set of the form $\{(1-t)a + tb; t \in \mathbb{R}\}$, where a and b are two distinct points in the plane. And a ***straight line segment*** (remember Euclid's *eutheĩa*) 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 $Z = \mathbb{Z}^2$ in the discussion that follows. The digitization of a straight line segment is the image under **dig** of $[a, b]$, thus

$$\mathbf{dig}([a,b]) = \bigcup_{t \in [0,1]} F(\{(1-t)a + tb\}) \subset \mathbb{Z}^2.$$

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 $\mathbf{Vo}_s(p) = \{x; \|x - p\|_\infty < \frac{1}{2}\}$, $p \in \mathbb{Z}^2$, then

$$F(\{x\}) = \left\{\left(\left\lfloor x_1 + \tfrac{1}{2}\right\rfloor, \left\lfloor x_1 + \tfrac{1}{2}\right\rfloor\right)\right\},$$

the unique point in $\mathbb{Z}^2$ 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 Rosenfeld's digitization 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, but this 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.

Definition

We shall say with Rosenfeld that a subset A of $\mathbb{R}^2$ 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.

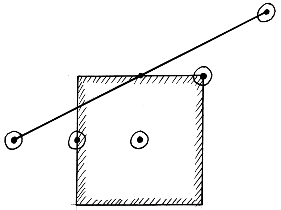
The theorems to be presented are due to Rosenfeld (1974) and give together a characterization of the digitization of a straight line segment. (The proof of the second theorem is new and is much shorter than the original proof.)

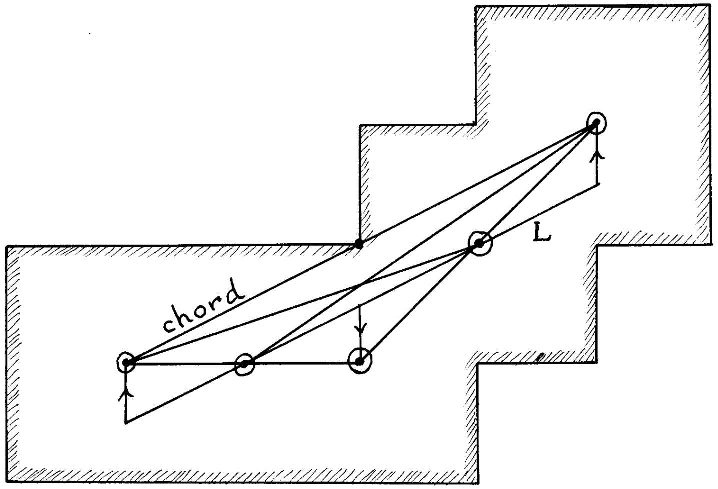
Theorem

The Rosenfeld digitization of a straight line segment has the chord property.

Example

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 $\left(0,-\frac{1}{2}\right)$ to be $(0,0)$, that of $\left(2,\frac{1}{2}\right)$ to be $(2,0)$, and that of $\left(4,1\frac{1}{2}\right)$ to be $(4,2)$. Then A is the digitization of the straight line segment $\left[\left(0,-\frac{1}{2}\right),\left(4,1\frac{1}{2}\right)\right]$.





The digitization just described 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 $\left(0,-\frac{1}{2}\right)$ and $\left(4,1\frac{1}{2}\right)$ and down by one half from $\left(2,\frac{1}{2}\right)$. Rosenfeld avoided this by always moving down in the case of half-integers.

Proof. 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 $D \subset \mathbb{Z}^2$ its digitization. Let p,q be two points in D, and y an arbitrary point on the segment $[p,q]$. We shall prove that there exists a point $d \in D$ such that $\|d-y\|_\infty < 1$.

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.

We shall now use only the vertical part of the cell,

$$C_{R,v}(p) = \left\{x; x_1 = p_1 \text{ and } p_2 - \tfrac{1}{2} < x_2 \leqslant p_2 + \tfrac{1}{2}\right\}.$$

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 $y_1 \in \mathbb{Z}$. In this case we define $s \in \mathbb{R}^2$ as the point in L with $s_1 = y_1$. The digitization $d = \mathbf{dig}(\{s\}) \in D$ of s satisfies $\|d - s\|_\infty \leqslant \frac{1}{2}$, thus $\|d - y\|_\infty \leqslant \|d - s\|_\infty + \|s - y\|_\infty \leqslant 1$. But can equality occur here? No. If we analyze the definition of the digitization we find that

$$y_2 - \tfrac{1}{2} < s_2 \leqslant y_2 + \tfrac{1}{2}$$

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

$$d_2 - 1 < s_2 - \tfrac{1}{2} \leqslant y_2 < s_2 + \tfrac{1}{2} \leqslant d_2 + 1,$$

so that actually $|d_2 - y_2| < 1$, while $|d_1 - y_1| = 0$, thus $y \in B_<(d, 1)$.

Next we consider the case $y_1 \notin \mathbb{Z}$; $m < y_1 < m+1$ for some integer m. We now define s, s' and s'' as the points on L such that $s_1 = y_1$, $s'_1 = m$ 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. We see that

$$d'_2 - 1 < s_2 - \tfrac{1}{2} \leqslant y_2 < s_2 + \tfrac{1}{2} \leqslant d'_2 + 1,$$

and conclude that $y \in B_<(d', 1) \cap B_<(d'', 1)$.

In case $d''_2 = d'_2 + 1$ we must have $y_2 > d'_2 - 1$ and $y_2 < d''_2 + 1$ so that $y \in B_<(d', 1) \cup B_<(d'', 1)$. 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 $\mathbb{Z}^2$ are ***eight-neighbors*** if their l^∞ distance is 1. Then a ***digital arc*** is a mapping from a finite integer interval $[a, b]_\mathbb{Z}$ into the plane $\mathbb{Z}^2$ 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

If a digital arc D in $\mathbb{Z}^2$ has the chord property, then it is the Rosenfeld digitization of some straight line segment in $\mathbb{R}^2$.

Lemma

Denote by $\pi_j \colon \mathbb{Z}^2 \to \mathbb{Z}$ 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 $\pi_j\big|_D \colon D \to \mathbb{Z}$, $j = 1, 2$, 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 the theorem. 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

$$S(\alpha,\beta,\gamma)=\{x\in\mathbb{R}^2;\alpha x_1+\beta\leqslant x_2\leqslant\alpha x_1+\gamma\}.$$

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 y be the point on $[p,q]$ with abscissa equal to that of s. (We note that p,s,q belong to $\mathbb{Z}^2$, while y 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 $s_2\geqslant y_2+1$, showing that y 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 $y\in B_<(d,1)$. So far we only know that y 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 $|d_1-y_1|=|d_1-s_1|\geqslant 1$, so that $\|y-d\|_\infty\geqslant|y_1-d_1|\geqslant 1$. 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. □

Convex functions on discrete sets with integer values

In Euclidean geometry, convex sets play an important role, and convex functions of real variables are of importance in several branches of mathematics, especially in optimization.

We will now propose definitions of convex sets and convex functions in a digital setting, definitions that have many desirable properties. They are in fact very simple—some may call them naive—but it seems to be necessary to investigate them first before one can go on to more sophisticated definitions. We shall show that functions which are both convex and concave have interesting relations to a refined definition of digital hyperplanes.

Since Rosenfeld's seminal paper (1974), where he explained how to digitize a real straight line segment, variants of this digitization have been introduced, among them digitizations which respect the Khalimsky topology; see Melin (2003). Here we shall not consider the Khalimsky topology, however. Instead, we shall look at definitions of digital hyperplanes, in particular that of Reveillès (1991), and compare them with the notion of digitally convex and concave functions.

Eckhardt studies no less than five different notions of convexity; one of them he calls H-convexity (2001:218)—this is the notion studied in the present paper.

When defining functions with integer values, we shall often use the *floor* and *ceiling functions* $\mathbb{R} \ni t \mapsto \lfloor t \rfloor, \lceil t \rceil \in \mathbb{Z}$. They are uniquely determined by the requirement that $\lfloor t \rfloor$ and $\lceil t \rceil$ be integers for every real number t and by the inequalities

$$t-1 < \lfloor t \rfloor \leqslant t < \lfloor t \rfloor + 1; \qquad \lceil t \rceil - 1 < t \leqslant \lceil t \rceil < t+1, \qquad t \in \mathbb{R}.$$

The real case

Let E be a vector space over $\mathbb{R}$. A subset A of E is said to be *convex* if the segment $[a,b] = \{(1-t)a + tb; 0 \leqslant t \leqslant 1\}$ is contained in A for every choice of $a, b \in A$; in other words if $\{a,b\} \subset A$ implies $[a,b] \subset A$.

And convex functions are most conveniently defined in terms of convex sets: a function $u\colon E \to [-\infty, +\infty] = \mathbb{R} \cup \{+\infty, -\infty\}$ is said to be *convex* if its *epigraph*

$$\operatorname{epi} u = \{(x,t) \in E \times \mathbb{R}; u(x) \leqslant t\}$$

is a convex set in $E \times \mathbb{R}$. For functions $f\colon P \to [-\infty, +\infty]_{\mathbb{Z}} = \mathbb{Z} \cup \{+\infty, -\infty\}$, where P is a subset of E, we define the epigraph as a subset of $P \times \mathbb{Z}$:

$$\operatorname{epi} f = \{(p,q) \in P \times \mathbb{Z}; f(p) \leqslant q\}.$$

It is also possible to go in the other direction and define convex sets in terms of convex functions: a set A in E is convex if and only if its indicator function $\mathbf{ind}_A$ is convex, where we define $\mathbf{ind}_A(x) = 0$ if $x \in A$ and $\mathbf{ind}_A(x) = +\infty$ otherwise. Naturally we would like to keep these equivalences in the digital case.

Important properties of the family of convex sets in a vector space are the following.

Proposition

If C_j, $j \in J$, are convex sets, then the intersection $\bigcap C_j$ is convex. If the index set J is ordered and filtering to the right, and if $(C_j)_{j \in J}$ is an increasing family of convex sets, then its union $\bigcup C_j$ is convex.

Because of this result, the intersection

$$\operatorname{cvx} A = \bigcap\big(C \in \mathscr{P}(E); C \text{ is convex and } C \supset A\big), \qquad A \in \mathscr{P}(E),$$

of all convex sets containing a given subset A of E is itself convex; it is called the *convex hull of A*.

Proposition

If u_j, $j \in J$, are convex functions on a vector space, then $\sup u_j$ is convex. If the index set J is ordered and filtering to the right, and if $(u_j)_{j\in J}$ is a decreasing family of convex functions, then its infimum $\inf u_j$ is convex.

To a given function $u\colon E \to [-\infty, +\infty]$ we associate two convex functions, viz. the supremum v of all convex minorants of u and the supremum w of all affine minorants of u. These functions are themselves convex, and of course $w \leqslant v \leqslant u$. We shall denote v by $\operatorname{cvx}(u)$, and w by $\tilde{\tilde{u}}$, a notation which will become clear when we have introduced the Fenchel transformation.

The function $v = \operatorname{cvx}(u)$ will be called the *convex hull of u*. For functions $f\colon P \to [-\infty, +\infty]$, P being a subset of E, we shall use the same notation. Such a function can be extended to a function u defined in all of E simply by taking $u = +\infty$ in the complement of P (then u and f have the same epigraph), and we define $\operatorname{cvx}(f) = \operatorname{cvx}(u)$.

In many cases, but not always, $\tilde{\tilde{u}}$ is equal to $\operatorname{cvx}(u)$. We note that $\tilde{\tilde{u}}$ has two extra properties in addition to being convex, properties that are not always shared by $\operatorname{cvx}(u)$. The first is that $\tilde{\tilde{u}}$ is lower semicontinuous for any topology for which the affine functions are continuous. The second is that if u takes the value $-\infty$ at a point, then $\tilde{\tilde{u}}$ must be identically equal to $-\infty$ (there are no affine minorants), whereas $\operatorname{cvx}(u)$ may take also finite values or $+\infty$.

We thus have

$$w = \tilde{\tilde{u}} \leqslant v = \operatorname{cvx}(u) \leqslant u.$$

However, in our research it will not be enough to study these functions: it is necessary to look at their epigraphs.

The epigraph $\operatorname{epi} u$ of u is a subset of $E \times \mathbb{R}$ and its convex hull $C = \operatorname{cvx}(\operatorname{epi} u)$ is easily seen to have the property

$$\text{(PLUS)} \qquad (x,s) \in C, s \leqslant t \text{ implies } (x,t) \in C.$$

The function $V_C(x) = \inf\big(t; (x,t) \in C\big)$ satisfies

$$\operatorname{epi}_{\mathrm{s}} V_C \subset \operatorname{cvx}(\operatorname{epi} u) \subset \operatorname{epi} V_C.$$

It is clear that V_C is convex and equal to the largest convex minorant $v = \operatorname{cvx}(u)$ of u already introduced. Thus $\operatorname{cvx}(u)$ can be retrieved from $\operatorname{cvx}(\operatorname{epi} u)$ but not conversely. We combine the results:

$$\operatorname{epi}_{\mathrm{s}} u \subset \operatorname{epi}_{\mathrm{s}}(\operatorname{cvx}(u)) \subset \operatorname{epi}_{\mathrm{s}}(\operatorname{cvx}(u)) \cup \operatorname{epi} u$$

$$\subset \operatorname{cvx}(\operatorname{epi} u) \subset \operatorname{epi}(\operatorname{cvx}(u)) \subset \operatorname{epi} \tilde{\tilde{u}},$$

and in general we cannot claim that $\operatorname{cvx}(\operatorname{epi} u)$ is an epigraph.

Convex sets which are squeezed in between the epigraph and the strict epigraph of a function will now play an important role. Such sets C satisfy $\operatorname{epi}_{\mathrm{s}} u \subset C \subset \operatorname{epi} u$ for some function u. This means that C is obtained from the strict epigraph by adding some points in the graph:

$$C = \operatorname{epi}_{\mathrm{s}} u \cup \{(x, u(x)); x \in A\} \subset \operatorname{epi}_{\mathrm{s}} u \cup \operatorname{graph} u = \operatorname{epi} u.$$

Extreme examples are the following. If u is strictly convex, like $u(x) = \|x\|_2^p$, $x \in \mathbb{R}^n$, with $1 < p < +\infty$, then any such set is convex, even though A may be very irregular. (For the Euclidean norm, a set of the form $B_<(c,r) \cup A$ is convex for any subset A of the sphere.) If on the other hand $u = 0$, then such a set is convex if and only if A itself is convex.

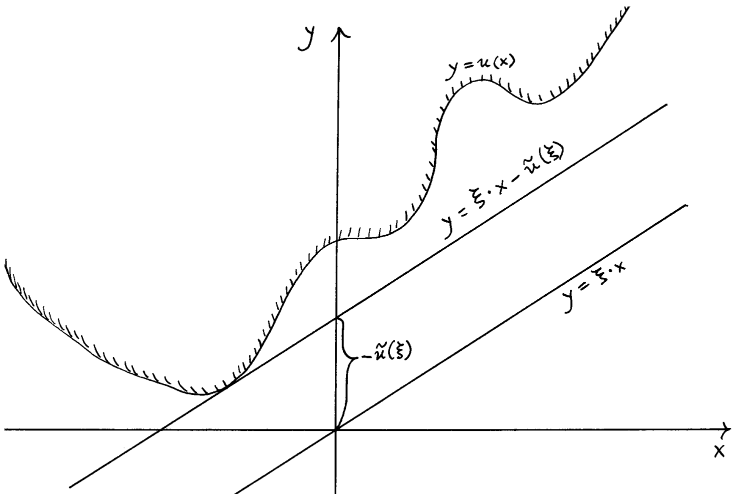
Definition

Let E be a real vector space and denote by $E^\star$ its algebraic dual (the set of all real-valued linear forms on E). For any function $u\colon E \to [-\infty, +\infty]$ we define its ***Fenchel transform*** $\widetilde{u}$ by

$$\widetilde{u}(\xi) = \sup_{x\in E} \bigl(\xi(x) - u(x)\bigr), \qquad \xi \in E^\star.$$

For any function $v\colon F \to [-\infty, +\infty]$ defined on a vector subspace F of $E^\star$ we define its ***Fenchel transform*** by

$$\widetilde{v}(x) = \sup_{\xi\in F} \bigl(\xi(x) - v(\xi)\bigr), \qquad x \in E.$$



The Fenchel transform of a function $u\colon \mathbb{R} \to \mathbb{R}$.

The second Fenchel transform $\tilde{\tilde{u}}$ of u is well-defined if we fix a subspace F of $E^\star$. This subspace can be anything between $\{0\}$ and all of $E^\star$, in particular we can take F as the topological dual E' of E if E is equipped with a vector space topology.

The restriction $\widetilde{u}\big|_F$ of the Fenchel transform to a subspace F of $E^\star$ describes all affine minorants of u with linear part in F: a pair $(\xi, \beta) \in F \times \mathbb{R}$ belongs to $\operatorname{epi} \widetilde{u}$ if and only if $x \mapsto \xi(x) - \beta$ is a minorant of u. This implies that $\tilde{\tilde{u}}$ is the supremum of all affine minorants of u with linear part in F. This function is a convex minorant of u, but it has the additional properties that it cannot take the value $-\infty$ unless it is the constant $-\infty$, and it is lower semicontinuous with respect to the topology $\sigma(E, F)$, the weakest topology on E for which all linear forms in F are continuous. One can prove that $\tilde{\tilde{u}}$ is the largest convex minorant of u with these properties.

Convex sets

Definition

Let E be a real vector space and fix a subset P of E. A subset A of P is said to be P-*convex* if there exists a convex set C in E such that $A = C \cap P$.

We are mostly interested in the case $E = \mathbb{R}^n$, $P = \mathbb{Z}^n$.

For digitizations of convex sets the mapping $C \mapsto C \cap \mathbb{Z}^n$ is not always satisfactory, because it yields the empty set for some long and narrow convex sets C. One might then want to replace it by a mapping like $C \mapsto (C + B) \cap \mathbb{Z}^n$, where B is some fixed set which guarantees that the image is nonempty when C is nonempty, e.g., $B = B_{\leqslant}(0, r)$, where $r = 1/2$ if we use the l^∞ norm in $\mathbb{R}^n$, $r = \sqrt{n}/2$ if we use the l^2 norm, or $r = n/2$ if we use the l^1 norm. However, for our purpose, when we apply this operation to the epigraph of a function, this phenomenon will not appear: the epigraph of a function with finite values always intersects $\mathbb{Z}^n \times \mathbb{Z}$ in a nonempty set.

Lemma

Given a vector space E and a subset P of E, the following properties are equivalent for any subset A of P.
1. A is P-convex;
2. $A = (\mathrm{cvx} A) \cap P$;
3. $A \supset (\mathrm{cvx} A) \cap P$.
4. For all n, all $a_0, \dots, a_n \in A$, and for all nonnegative numbers $\lambda_0, \dots, \lambda_n$ with $\sum_0^n \lambda_j = 1$, if $\sum_0^n \lambda_j a_j \in P$, then $\sum_0^n \lambda_j a_j \in A$.

Definition

Fix two subsets P and Q of a vector space E and define an operator $\gamma = \gamma_{P,Q} \colon \mathscr{P}(E) \to \mathscr{P}(P)$ by $\gamma(A) = \mathrm{cvx}(A \cap Q) \cap P$.

We can think of $E = \mathbb{R}^n$, $P = m\mathbb{Z}^n$, $m = 1, 2, \dots,$ and $Q = \mathbb{Z}^n$. We note that $\gamma(C)$ is P-convex if C is convex in $\mathbb{R}^n$.

Lemma

The mapping γ is increasing; it satisfies $\gamma(\gamma(A)) \subset \gamma(A)$; and it satisfies $A \subset \gamma(A)$ if $A \subset P \cap Q$. Thus $\gamma\big|_{\mathscr{P}(P)}$ is a closure operator in $\mathscr{P}(P)$ if $Q \supset P$.

Proof. The mapping $\gamma = j_P \circ \mathrm{cvx} \circ j_Q$ is a composition of three increasing mappings, viz. j_Q (intersection with Q), cvx (taking the convex hull), and j_P (intersection with P), and as such itself increasing. The composition $\gamma \circ \gamma$ is equal to $j_P \circ \mathrm{cvx} \circ j_Q \circ j_P \circ \mathrm{cvx} \circ j_Q$, which is smaller than $j_P \circ \mathrm{cvx} \circ \mathrm{cvx} \circ j_Q = j_P \circ \mathrm{cvx} \circ j_Q = \gamma$. Finally, it is clear that $\gamma(A)$ contains A if A is contained in $P \cap Q$. If $Q \supset P$, then γ is increasing, idempotent and extensive, thus a closure operator in $\mathscr{P}(P)$. □

Proposition

Let E be a real vector space and P any subset of E. Then A is P-convex iff $A = \gamma(A)$ for all $Q \supset P$ iff $A = \gamma(A)$ for some $Q \supset P$.

Corollary

If $A = C \cap P$ for some convex set $C \subset E$, then $C \supset \gamma(A)$ for any Q.

Thus in the definition of P-convex sets we may always take $C = \gamma(A) = \mathrm{cvx} A$ provided $Q \supset P$.

It is now easy to prove the following result.

Proposition

Let E be a vector space and P any subset of E. If A_j, $j \in J$, are P-convex sets, then the intersection $\bigcap A_j$ is P-convex. If the index set J is ordered and filtering to the right, and if $(A_j)_{j \in J}$ is an increasing family of P-convex sets, then its union $\bigcup A_j$ is also P-convex.

While the intersection of two P-convex epigraphs gives a reasonable result, the intersection of an epigraph and a hypograph may consist of two points quite far from each other:

Example

Let $A = \{p \in \mathbb{Z}^2; p_2 \geqslant p_1/m\}$ and $B = \{p \in \mathbb{Z}^2; p_2 \leqslant p_1/m\}$, where $m \in \mathbb{N} \smallsetminus \{0\}$. Then A and B are $\mathbb{Z}^2$-convex and their intersection consists of all points (mp_2, p_2), $p_2 \in \mathbb{Z}$. We can easily modify the example so that the intersection consists of exactly two points, $(0,0)$ and $(m,1)$, where m is as large as we please.

Convex functions with integer values

Definition

Let E be a vector space and P any of its subsets. A function $f\colon P \to [-\infty, +\infty]_{\mathbb{Z}}$ is said to be $(P \times \mathbb{Z})$-convex if its epigraph

$$\operatorname{epi} f = \{(p,t) \in P \times \mathbb{Z}; f(p) \leqslant t\}$$

is a $(P \times \mathbb{Z})$-convex subset of $E \times \mathbb{R}$.

We have mainly the case $E = \mathbb{R}^n$ and $P = \mathbb{Z}^n$ in mind.

If $f\colon P \to [-\infty, +\infty]_{\mathbb{Z}}$ is a P-convex function, then there is a convex set C in $E \times \mathbb{R}$ such that $C \cap (P \times \mathbb{Z}) = \operatorname{epi} f$. In view of the previous corollary, the smallest such set C is the convex hull of $\operatorname{epi} f$. However, a set C such that $C \cap (P \times \mathbb{Z}) = \operatorname{epi} f$ does not necessarily have the property (PLUS), so we introduce

$$C^+ = \{(x,t) \in E \times \mathbb{R}; \exists s \leqslant t \text{ with } (x,s) \in C\}.$$

There is a function $V_{C^+}\colon E \to [-\infty, +\infty]$ such that

$$\operatorname{epi}_s V_{C^+} \subset C^+ \subset \operatorname{epi} V_{C^+}.$$

It would perhaps seem natural to require that C^+ be closed or open so that one could always take either the epigraph or the strict epigraph of V_{C^+}, but simple examples will show that this is not possible. We note that when we take $C = \operatorname{cvx}(\operatorname{epi} f)$, then $C^+ = C$.

Some care is needed, because even if $\operatorname{epi} f$ is closed, its convex hull need not be closed:

Example

Let $f_0(p) = \lceil \alpha p \rceil$, $p \in \mathbb{Z}$, where α is irrational. We also define $f_1(p) = f_0(p)$ for $p \in \mathbb{Z} \smallsetminus \{0\}$ and $f_1(0) = 1$. These functions are easily seen to be $(\mathbb{Z} \times \mathbb{Z})$-convex. Indeed, $\operatorname{cvx}(\operatorname{epi} f_1)$ is the open half plane $C_1 = \{(x,t); t > \alpha x\}$, a strict epigraph, and $\operatorname{cvx}(\operatorname{epi} f_0)$ is the convex set $C_0 = C_1 \cup \{(0,0)\}$, which is neither an epigraph nor a strict epigraph. (However, also the closed half plane $\{(x,t); t \geqslant \alpha x\}$ intersects $\mathbb{Z}^2$ in $\operatorname{epi} f_0$.) We finally note that the functions $-f_0$ and $-f_1$ are $(\mathbb{Z} \times \mathbb{Z})$-convex as well.

Proposition

Let $u\colon E \to [-\infty, +\infty]$ be a convex function on a vector space E. Let P be a subset of E. Then the restrictions $\lfloor u \rfloor\big|_P$ and $\lceil u \rceil\big|_P$ are $(P \times \mathbb{Z})$-convex. In particular $\lceil \operatorname{cvx} g \rceil\big|_P$ and $\lceil \tilde{\tilde{g}} \rceil\Big|_P$ are $(P \times \mathbb{Z})$-convex for any function $g\colon P \to [-\infty, +\infty]_{\mathbb{Z}}$.

Proof. Writing $f = \lfloor u \rfloor\big|_P$ and $g = \lceil u \rceil\big|_P$ we have

$$u - 1 < f \leqslant u \text{ and } u \leqslant g < u + 1 \text{ in } P,$$

which implies that $\operatorname{epi}_{\mathrm{s}}(u-1) \cap (P \times \mathbb{Z}) = \operatorname{epi} f$ and $\operatorname{epi} u \cap (P \times \mathbb{Z}) = \operatorname{epi} g$. Hence the functions f and g are $(P \times \mathbb{Z})$-convex. □

Theorem

Let E be a vector space and P one of its subsets. For any $(P \times \mathbb{Z})$-convex function $f\colon P \to \mathbb{Z}$ we have $\operatorname{cvx} f \leqslant \lceil \operatorname{cvx} f \rceil \leqslant f \leqslant \operatorname{cvx} f + 1$ in P.

The first two inequalities are easy; the third is the essential result of the theorem.

We define

$$P^j = \{p \in P; f(p) = \lceil (\operatorname{cvx} f)(p) \rceil + j\}, \qquad j = 0, 1.$$

In view of the last theorem we have $P = P^0 \cup P^1$. We also define

$$A^j = \{p \in P; f(p) = (\operatorname{cvx} f)(p) + j\}, \qquad j = 0, 1.$$

Corollary

With f as in the theorem, P can be divided into three disjoint sets: $P^0 \smallsetminus A^0$, A^0, and $A^1 = P^1$. The first set is precisely the set of points p such that $(\operatorname{cvx} f)(p)$ is not an integer.

Proof. It is clear that the three sets $P^0 \smallsetminus A^0$, A^0 and P_1 are pairwise disjoint. It is also easy to see that $p \in A^0 \cup A^1$ if and only if $(\operatorname{cvx} f)(p)$ is an integer. It follows that $A^j \subset P^j$. Finally, we shall prove that $P^1 \subset A^1$. If $p \in P^1$, then $\lceil (\operatorname{cvx} f)(p) \rceil$ is equal to $f(p) - 1$. But we always have $(\operatorname{cvx} f)(p) \geqslant f(p) - 1$, so that $(\operatorname{cvx} f)(p) = \lceil (\operatorname{cvx} f)(p) \rceil$ and p belongs to A^1. □

Let us say that a function $u\colon \mathbb{R}^n \to [-\infty,+\infty]$ is *of fast growth* if for any constant c the set $\{x \in \mathbb{R}^n; u(x) \leqslant c\|x\|_2\}$ is bounded. The same terminology applies to a function defined in a subset P of $\mathbb{R}^n$; we understand that it takes the value $+\infty$ outside P. In particular, if f is equal to plus infinity outside a bounded set, it is of fast growth.

Theorem

Let P be a discrete subset of $\mathbb{R}^n$ and let $f\colon P \to [-\infty,+\infty]_{\mathbb{Z}}$ be a function of fast growth. Then f is $(P\times\mathbb{Z})$-convex if and only if $f = \lceil \mathrm{cvx} f \rceil$, in other words the set P^1 is empty. We have for all $(P\times\mathbb{Z})$-convex fucntions f,

$$(\mathrm{cvx} f)(p) \leqslant f(p) < (\mathrm{cvx} f)(p) + 1, \qquad p \in P.$$

It is equivalent to say that there exists a convex function $u\colon \mathbb{R}^n \to [-\infty,+\infty]$ such that $f = \lceil u \rceil$.

Proposition

Let E be a vector space and P any of its subsets. If f_j, $j \in J$, are $(P\times\mathbb{Z})$-convex functions, then $\sup f_j$ is $(P\times\mathbb{Z})$-convex. If the index set J is ordered and filtering to the right, and if $(f_j)_{j\in J}$ is a decreasing family of $(P\times\mathbb{Z})$-convex functions, then its infimum $\inf f_j$ is $(P\times\mathbb{Z})$-convex as well.

Affine functions

A function u such that $-u$ is convex is called ***concave***.

A real-valued function on $\mathbb{R}^n$ which is both convex and concave is necessarily ***affine***, i.e., of the form $u(x) = \alpha \cdot x + \beta$ for some $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$.

We shall now investigate in the discrete case functions which are both convex and concave.

Proposition

Let P be a nonempty subset of a vector space E and $f\colon P \to \mathbb{R}$ a real-valued function. Given a linear form $\alpha \in E^\star$ and a real number β we let $h_{\alpha,\beta}$ be the smallest constant $h \in [0,+\infty]$ such that

$$0 \leqslant \alpha(p) + \beta \leqslant f(p) \leqslant \alpha(p) + \beta + h, \qquad p \in P.$$

We let $h_\alpha = \inf_{\beta\in\mathbb{R}} h_{\alpha,\beta}$ be the smallest constant h such that this holds for some $\beta \in \mathbb{R}$. Then $h_\alpha = \widetilde{f}(\alpha) + \widetilde{g}(-\alpha)$, where for ease in notation we have written g for $-f$. Moreover, $h_\alpha = h_{\alpha,\beta}$ for a unique β, viz. $\beta = -\widetilde{f}(\alpha)$.

Proof. The inequality $\alpha(p)+\beta \leqslant f(p)$ for all $p \in P$ is equivalent to $\widetilde{f}(\alpha) \leqslant -\beta$, and the inequality $f(p) = -g(p) \leqslant \alpha(p)+\beta+h$ for all $p \in P$ is equivalent to $\widetilde{g}(-\alpha) \leqslant \beta+h$. Therefore $\alpha+\beta \leqslant f \leqslant \alpha+\beta+h$ implies that $\widetilde{f}(\alpha)+\widetilde{g}(-\alpha) \leqslant -\beta+(\beta+h) = h$.

Conversely, if h is a real number and $\widetilde{f}(\alpha)+\widetilde{g}(-\alpha) \leqslant h$, then $\widetilde{f}(\alpha)$ is a real number: $\widetilde{f}(\alpha) = -\infty$ would imply that f is identically equal to $+\infty$, which is excluded by hypothesis, and $\widetilde{g}(-\alpha) = -\infty$ would imply that f is identically $-\infty$, which is also excluded by hypothesis; finally, the inequality excludes that $\widetilde{f}(\alpha)$ is equal to $+\infty$. Therefore $\beta = -\widetilde{f}(\alpha)$ (obviously the best choice of β) yields $\widetilde{f}(\alpha) \leqslant -\beta$ and $\widetilde{g}(-\alpha) \leqslant \beta+h$.

The infimum of all real h satisfying $\alpha+\beta \leqslant f \leqslant \alpha+\beta+h$ is equal to the infimum of all real h satisfying $\widetilde{f}(\alpha)+\widetilde{g}(-\alpha) \leqslant h$, which completes the proof. □

Proposition

Let E be a vector space and P a subset such that $\operatorname{cvx} P = E$. Let a real-valued function $f\colon P \to \mathbb{R}$ be given, and let $h_* = \inf_{\alpha \in E^\star} h_\alpha$ be the smallest constant h such that the double inequality $\alpha+\beta \leqslant f \leqslant \alpha+\beta+h$ holds for some $\alpha \in E^\star$ and some $\beta \in \mathbb{R}$. Assume that h_* is finite. Then $\operatorname{cvx} f + \operatorname{cvx}(-f)$ is constant and equal to $-h_*$.

Proof. Let h be a number such that $\alpha+\beta \leqslant f \leqslant \alpha+\beta+h$ in P for some $\alpha \in E^\star$ and some $\beta \in \mathbb{R}$. Then

$$\alpha+\beta \leqslant u \leqslant f \leqslant -v \leqslant \alpha+\beta+h \text{ in } P,$$

where $u = \operatorname{cvx} f$ and $v = \operatorname{cvx}(-f)$. Adding v to all members we obtain

$$\alpha+\beta+v \leqslant u+v \leqslant f+v \leqslant 0 \leqslant \alpha+\beta+h+v \text{ in } P.$$

We see that $u+v$ is a convex function which is nonpositive in all of P, thus also in $\operatorname{cvx} P$, which by hypothesis is equal to E.

But such a function must be constant; let us define $\omega = -(u+v) \geqslant 0$. By the same argument, $v+\alpha$ is a constant γ.

We now have $\gamma+\beta \leqslant -\omega \leqslant 0 \leqslant \gamma+\beta+h$, which shows that $h \geqslant \omega$, and, by taking the infimum over all such h, that $h_* \geqslant \omega$.

Conversely, we note that $-\omega \leqslant f+\gamma-\alpha \leqslant 0$, thus $\alpha-\gamma-\omega \leqslant f \leqslant \alpha-\gamma$, which shows that $\omega \geqslant h_\alpha \geqslant h_*$. We conclude that $\omega = h_*$. □

Theorem

Let E be a vector space and P a subset of E such that $\operatorname{cvx} P = E$. If both functions $f\colon P \to \mathbb{Z}$ and $-f$ are $(P \times \mathbb{Z})$-convex, then f deviates at most by $\frac{1}{2}$ from an affine function: there exist a linear form $\alpha \in E^\star$ and constants $\beta, \omega \in \mathbb{R}$ such that

$$0 \leqslant f(p)-\alpha(p)-\beta \leqslant \omega \leqslant 1, \qquad p \in P.$$

The best constant ω is equal to the constant $-\operatorname{cvx} f - \operatorname{cvx}(-f)$. Also $(\operatorname{cvx} f)(x) = \alpha(x)+\beta$ and $\operatorname{cvx}(-f)(x) = -\alpha(x)-\beta-\omega$ if ω is chosen as the smallest possible constant.

We rewrite the theorem in the most common situation:

Corollary

If both $f\colon \mathbb{Z}^n \to \mathbb{Z}$ and $-f$ are $(\mathbb{Z}^n \times \mathbb{Z})$-convex, then there exist $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ such that

$$0 \leqslant f(p) - \alpha \cdot p - \beta \leqslant \omega, \qquad p \in \mathbb{Z}^n,$$

where ω is the constant $-\mathrm{cvx} f - \mathrm{cvx}(-f) \leqslant 1$.

Is it possible to take one of the inequalities here strict? Actually not.

Examples show that there is a choice between the intervals $[0, \omega[$ and $]0, \omega]$ in the inequality for different values of p. This choice is made precise in the following result.

Theorem

Let $f\colon \mathbb{Z}^n \to \mathbb{Z}$ and $-f$ be $(\mathbb{Z}^n \times \mathbb{Z})$-convex and let $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ be such that $0 \leqslant f(p) - \alpha \cdot p - \beta \leqslant \omega$ holds with $\omega = h_*$, i.e., with the smallest h possible. Define

$$D^j = \{(p, f(p)) \in \mathbb{Z}^n \times \mathbb{Z}; f(p) = \alpha \cdot p + \beta + j\omega\}, \qquad j = 0, 1,$$

and

$$A^j = \pi_{n+1}(D^j) = \{p \in \mathbb{Z}^n; f(p) = \alpha \cdot p + \beta + j\omega\}, \qquad j = 0, 1,$$

where $\pi_{n+1}\colon \mathbb{Z}^n \times \mathbb{Z} \to \mathbb{Z}^n$ denotes the projection which forgets the last coordinate. Assume that $\omega > 0$. Then A^0 and A^1 are disjoint, and D^0 and D^1 are $(\mathbb{Z}^n \times \mathbb{Z})$-convex.

Digital hyperplanes

The concept of ***naive discrete line*** was introduced by Reveillès (1991:48). Such a line is defined to be the set of all integer points $p \in \mathbb{Z}^2$ such that $0 \leqslant \alpha_1 p_1 + \alpha_2 p_2 + \beta < \max(|\alpha_1|, |\alpha_2|)$ for some β, where α_1 and α_2 are relatively prime integers.

Generalizing this slightly, we define a ***naive digital hyperplane*** as the set of all points $p \in \mathbb{Z}^n$ which satisfy the double inequality

$$0 \leqslant \alpha \cdot p + \beta < h,$$

for some $\alpha \in \mathbb{R}^n \smallsetminus \{0\}$ and some $\beta \in \mathbb{R}$, where $h = \|\alpha\|_\infty$.

We remark that one can always interchange the strict and the non-strict inequalities: the set just defined can equally well be defined by

$$0 < (-\alpha) \cdot p - \beta + h \leqslant h.$$

The precise size of h is important for the representation of the hyperplane as the graph of a function of $n - 1$ variables as shown by the following result.

Theorem

Define

$$T=\{p\in\mathbb{Z}^n; 0\leqslant \alpha\cdot p+\beta\leqslant h\} \text{ and } T_{\mathrm{s}}=\{p\in\mathbb{Z}^n; 0<\alpha\cdot p+\beta<h\},$$

where $\alpha\in\mathbb{R}^n\smallsetminus\{0\}$, $\beta\in\mathbb{R}$ *and* $h>0$, *and let*

$$T^j=\{p\in\mathbb{Z}^n;\alpha\cdot p+\beta=jh\},\qquad j=0,1.$$

Let D *be a subset of* $\mathbb{Z}^n$ *which is contained in* T *and contains* T_{s} *and define* $D_{\mathrm{s}}=D\cap T_{\mathrm{s}}$ *and* $D^j=D\cap T^j$. *Fix an integer* $k=1,\ldots,n$ *and let* $\pi_k\colon\mathbb{Z}^n\to\mathbb{Z}^{n-1}$ *be the projection which forgets the* k^{th} *coordinate. Then* $\pi_k\big|_D$ *is injective if* $h<|\alpha_k|$, *and* $\pi_k\big|_D$ *is surjective if* $h>|\alpha_k|$. *If* $h=|\alpha_k|$, *then* $\pi_k\big|_D$ *is injective if and only if* $\pi_k(D^0)$ *and* $\pi_k(D^1)$ *are disjoint, and* $\pi_k\big|_D$ *is surjective if and only if* $\pi_k(D^0\cup D^1)=\pi_k(T^0\cup T^1)$.

Proof. For ease in notation we let $k=n$ and write $p'=(p_1,\ldots,p_{n-1})$ and similarly for α. Then p belongs to T if and only if

$$-\alpha'\cdot p'-\beta\leqslant\alpha_n p_n\leqslant-\alpha'\cdot p'-\beta+h,$$

and p belongs to T_{s} if and only if

$$-\alpha'\cdot p'-\beta<\alpha_n p_n<-\alpha'\cdot p'-\beta+h.$$

Clearly for every p' there is at most one p_n which satisfies the inequalities if $h<|\alpha_n|$ or if $h=|\alpha_n|$ and $(\alpha'\cdot p'+\beta)/h$ is not an integer. Also there is at least one p_n if $h>|\alpha_n|$ or if $h=|\alpha_n|$ and $(\alpha'\cdot p'+\beta)/h$ is not an integer. Here it does not matter whether we use the first or the second inequality, so the conclusion holds also for D.

The case when $h=|\alpha_n|$ and $(\alpha'\cdot p'+\beta)/h$ is an integer remains to be considered. Then we see that there are two values of p_n which satisfy the nonstric inequality and and none that satisfies the strict inequality. Hence there is at most one p_n such that (p',p_n) belongs to $D=D^0\cup D_{\mathrm{s}}\cup D^1$ if and only if $\pi_k(D^0)$ and $\pi_k(D^1)$ are disjoint. There is at least one p_n such that (p',p_n) belongs to D if and only if $\pi_k(D^0\cup D^1)$ contains every point in the projection of $T^0\cup T^1$. This completes the proof. □

We do not suppose here that $h=\|\alpha\|_\infty$. However, this is the most natural case: we then know that $\pi_k\big|_D$ is a bijection for any k such that $|\alpha_k|=\|\alpha\|_\infty$ and the conditions on the D^j are satisfied, and that $\pi_j\big|_D$ is surjective for all j such that $|\alpha_j|<\|\alpha\|_\infty$.

It seems reasonable to propose the following definition.

Definition

A ***refined digital hyperplane*** is a $\mathbb{Z}^n$-convex subset D of $\mathbb{Z}^n$ which is contained in the slab T and contains the strict slab T_{s} for some $\alpha\in\mathbb{R}^n\smallsetminus\{0\}$, $\beta\in\mathbb{R}$, and $h>0$; and in addition is such that, for at least one k such that $|\alpha_k|=h$, the sets $D^j=D\cap T^j$ have disjoint projections $\pi_k(D^j)$, and $\pi_k(D^0\cup D^1)=\pi_k(T^0\cup T^1)$.

The naive hyperplanes now appear as a special case, viz. when $D^0 = T^0$, and D^1 is empty, or conversely, and $|\alpha_k| = \|\alpha\|_\infty$.

Example

Define $D = (D^0 \times \{0\}) \cup (D^1 \times \{1\})$, where D^j, $j = 0, 1$, are two subsets of $\mathbb{Z}^{n-1}$ such that $D^1 = \mathbb{Z}^{n-1} \smallsetminus D^0$. Then D is a refined digital hyperplane if and only if both D^0 and D^1 are $\mathbb{Z}^{n-1}$-convex.

Example

Define $D = \{(p_1, p_1) \in \mathbb{Z}^2; p_1 \leqslant 0\} \cup \{(p_1, p_1 + 1) \in \mathbb{Z}^2; p_1 \geqslant 0\}$. This is a refined digital hyperplane with $|\alpha_1| = |\alpha_2| = \|\alpha\|_\infty = 1$. The projection π_1 satisfies the requirements in the definition, but π_2 does not.

The following result motivates the definition just given and relates it to the digitally convex functions we have introduced.

Theorem

A subset D of $\mathbb{Z}^n$ is a refined digital hyperplane if and only if, after a permutation of the coordinates, it is the graph of a function $f\colon \mathbb{Z}^{n-1} \to \mathbb{Z}$ such that both f and $-f$ are $(\mathbb{Z}^{n-1} \times \mathbb{Z})$-convex.

Proof. Let f be a $(\mathbb{Z}^{n-1} \times \mathbb{Z})$-convex function such that also $-f$ is $(\mathbb{Z}^{n-1} \times \mathbb{Z})$-convex. Then $D = \operatorname{graph} f$ is a refined digital hyperplane according to the previous theorem.

Conversely, if D is a refined digital hyperplane and $h = |\alpha_n|$, then the projection $\pi_n\big|_D$ is bijective, and this allows us to define a function $f\colon \mathbb{Z}^{n-1} \to \mathbb{Z}$, $f(p') = -\alpha' \cdot p' - \beta + jh$ with $j = 0$ or 1 being uniquely determined by the requirements on the D^j. This function as well as its negative are $(\mathbb{Z}^{n-1} \times \mathbb{Z})$-convex, since both its epigraph and its hypograph are $\mathbb{Z}^n$-convex. To wit, assuming α_n to be positive, its epigraph is equal to $D + (\{0\} \times \mathbb{N})$, and its hypograph is equal to $D + (\{0\} \times (-\mathbb{N}))$. □

Conclusion: Functions that are both convex and concave are of interest as candidates for defining digital hyperplanes; in fact we have shown that they define sets which are precisely the sets satisfying a refined definition of digital hyperplanes.

Remark: For other classes of convex functions, see my paper on discrete optimization.

Topology

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 $\varnothing$ and X are always open subsets of X.

A set F is called ***closed*** if its complement $X \smallsetminus F$ is open.

The ***closure*** of a subset A of a topological space X is the intersection of all closed sets containing A. It will be denoted by $\overline{A}$. It is the smallest closed set which contains A.

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 $\mathscr{U}_1(X) \subset \mathscr{U}_2(X)$.

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.

These two extreme topologies are not so interesting to work with.

In a metric space there is a topology defined by the metric: it consists of all unions of strict balls $B_<(c, r)$.

If we consider $\mathbb{Z}^2$ as a subspace of $\mathbb{R}^2$ and define a set $U \subset \mathbb{Z}^2$ to be open if there is an open set V in $\mathbb{R}^2$ such that $U = V \cap \mathbb{Z}^2$, then every subset of $\mathbb{Z}^2$ is open, i.e., $\mathbb{Z}^2$ gets the discrete topology. This is, as already noted, not interesting.

Instead we shall define another topology on $\mathbb{Z}^2$, called the Khalimsky topology, and which can be described as a quotient space of $\mathbb{R}^2$. So we shall think of $\mathbb{Z}^2$ not as a subspace of $\mathbb{R}^2$ but as a quotient space!

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***. Of the four, only three are different in the sense that they cannot be obtained from another one by renaming the points.

The Sierpiński topology gives a pretaste of the Khalimsky topology.

Continuous mappings

Let $f\colon X \to Y$ be a mapping of a topological space X into a topological space Y. We say that f is ***continuous*** if the preimage

$$f^*(V) = \{x \in X; f(x) \in V\}$$

is open in X for every open subset V of Y.

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. 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

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 its image $\operatorname{im} f$.

Corollary

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.

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.

Let $f\colon \mathbb{R} \to \mathbb{Z}$ be a surjective mapping. Then we can define a topology in $\mathbb{Z}$ by declaring a subset V of $\mathbb{Z}$ to be open if its preimage

$$f^*(V) = \{x \in \mathbb{R}; f(x) \in V\}$$

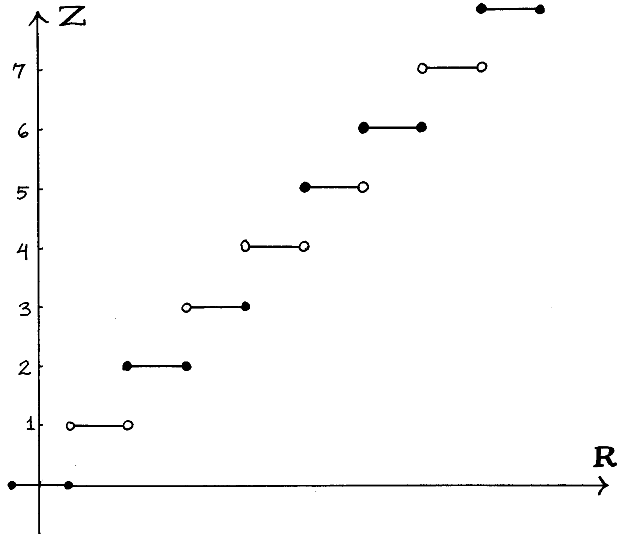
is open in $\mathbb{R}$. It is easy to prove that this defines a topology.

In this situation we say that $\mathbb{Z}$ is a quotient space of $\mathbb{R}$.

Now there exist very many surjective mappings $\mathbb{R} \to \mathbb{Z}$. It is not unnatural to restrict attention to increasing surjections $f\colon \mathbb{R} \to \mathbb{Z}$. Then $\{x; f(x) = n\}$ is an interval for every integer n; denote its endpoints by a_n and $b_n \geqslant a_n$, so that

$$]a_n, b_n[\subset \{x; f(x) = n\} \subset [a_n, b_n].$$

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 $\mathbb{Z}$. Then $f(x) = \lfloor x + \frac{1}{2} \rfloor$ for all $x \in \mathbb{R} \smallsetminus (\mathbb{Z} + \frac{1}{2})$, and $f(n + \frac{1}{2}) = n$ or $f(n + \frac{1}{2}) = n + 1$ for $n \in \mathbb{Z}$. The topology is therefore determined if we know for which n we have $f(n + \frac{1}{2}) = n$.



$\{1\}$, $\{4\}$ and $\{7\}$ are open sets; $\{2\}$, $\{6\}$ and $\{8\}$ are closed.

For every subset A of $\mathbb{Z}$ we get a topology on $\mathbb{Z}$ by declaring that $f(n + \frac{1}{2})$ shall be equal to n for $n \in A$ and that, for all other real numbers 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 . . .

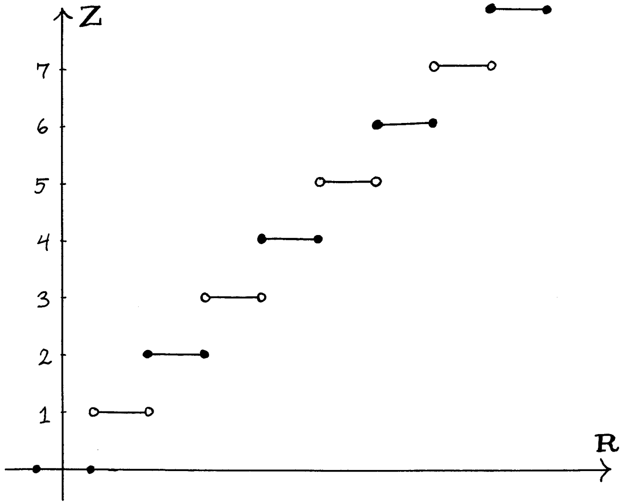
It is natural to think of $\mathbb{Z}$ as an approximation of the real line $\mathbb{R}$ and to consider mappings $f\colon \mathbb{R} \to \mathbb{Z}$ expressing this idea. We may define $f(x)$ to be the integer closest to x; this is well-defined unless x is a half-integer: $f(x) = \lfloor x + \frac{1}{2} \rfloor$ when $x \in \mathbb{R} \smallsetminus \left(\mathbb{Z} + \frac{1}{2}\right)$. 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 $\{x; f(x) = n\} = \left]n - \frac{1}{2}, n + \frac{1}{2}\right]$, the topology defined in the corollary is called the ***right topology*** on $\mathbb{Z}$; if we choose the second, we obtain the ***left topology*** on $\mathbb{Z}$; 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***; $\mathbb{Z}$ with this topology is called the ***Khalimsky line***.

Thus a set V of integers is open for the Khalimsky topology if and only if, for every even number $2k \in V$, also its odd neighbors $2k - 1$ and $2k + 1$ belong to V.

A set F of integers is closed for the Khalimsky topology if and only if for every odd number $2k + 1 \in F$, also its even neighbors $2k$ and $2k + 2$ belong to F.



$\{1\}$, $\{3\}$, $\{5\}$ and $\{7\}$ are open sets; $\{0\}$, $\{2\}$, $\{6\}$ and $\{8\}$ are closed.

The Khalimsky line is connected, but the complement of any point is disconnected.

And the real line $\mathbb{R}$ has the same property: $\mathbb{R}$ is connected and $\mathbb{R} \smallsetminus \{a\}$ is disconnected for any $a \in \mathbb{R}$.

Among all the topologies defined by increasing surjections $f\colon \mathbb{R} \to \mathbb{Z}$ 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.

Separation axioms

A ***neighborhood*** of a point $x \in X$ is a set V such that there is some open set U with $x \in U \subset V$.

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 $y \in \overline{\{x\}}$ (i.e., y is in the closure of the singleton set $\{x\}$). The relation $x \in N(y)$ defines a preorder in X. 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. 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.

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.

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***.

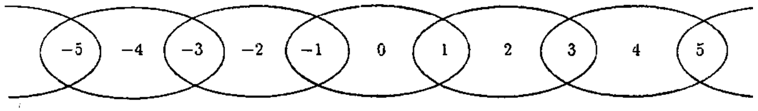
The intersection $N(x)$ of all neighborhoods of a point x is open for all x if and only if the space is a smallest-neighborhood space.

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 closed set of a smallest-neighborhood space satisfies the axioms of the open sets of a topology. We can declare a closed set to be open—in this way we get a new topology. 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.

We can define a topology on the digital line $\mathbb{Z}$ by declaring all odd points to be open, thus $N(2k+1) = \{2k+1\}$, and all even points to have a smallest neighborhood $N(2k) = \{2k-1, 2k, 2k+1\}$. 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.

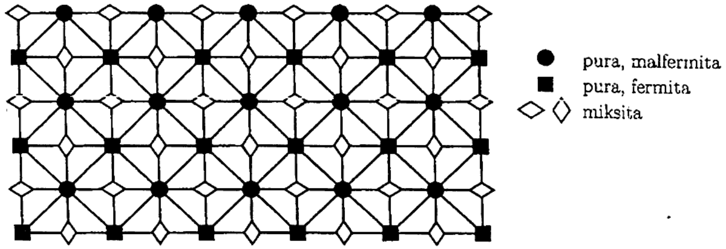


The Khalimsky line.

A ***Khalimsky interval*** is an interval $[a,b]_{\mathbb{Z}} = [a,b]_{\mathbb{R}} \cap \mathbb{Z}$ equipped with the topology induced by the Khalimsky topology on $\mathbb{Z}$. A ***Khalimsky circle*** is a quotient space $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$ of the Khalimsky line for some even integer $m \geqslant 4$. (If m is odd, the quotient space receives the chaotic topology.)

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

There are, however, other topologies in $\mathbb{Z}^2$ which are of interest: we may declare $\{x \in \mathbb{Z}^2; \|x - c\|_1 \leqslant 1\}$ to be open for any c such that $\sum c_j \in 2\mathbb{Z}$ as well as all intersections of such sets (Wyse et al. 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.



The Khalimsky plane.

To exhibit some of the analogies between topological spaces and preordered sets, let us list some properties of continuous and increasing mappings.

Mappings $X \longrightarrow Y$ between topological spaces	*Mappings $X \longrightarrow Y$ between preordered sets*
X has the discrete topology ⇒ all mappings are continuous	X has the discrete order ⇒ all mappings are increasing
Y has the chaotic topology ⇒ all mappings are continuous	Y has the chaotic preorder ⇒ all mappings are increasing
X has the chaotic topology and Y has a Kolmogorov topology ⇒ only the constants are continuous	X has the chaotic preorder and Y is ordered ⇒ only the constants are increasing
Y has the discrete topology and X is connected ⇒ only the constants are continuous	Y has the discrete order and X is connected ⇒ only the constants are increasing

The Khalimsky plane

We observe that the following functions are continuous:
(1) $\mathbb{Z} \ni x \mapsto a \in \mathbb{Z}$, where a is a constant;
(2) $\mathbb{Z} \ni x \mapsto \pm x + c \in \mathbb{Z}$, 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.

We can describe continuity in terms of even and odd coordinates. 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 $\mathbb{Z}$,

$$\cdots \preccurlyeq -2 \succcurlyeq -1 \preccurlyeq 0 \succcurlyeq 1 \preccurlyeq 2 \succcurlyeq 3 \preccurlyeq 4 \succcurlyeq \cdots$$

In $\mathbb{Z}^2$, $(0,0) \succcurlyeq (1,0), (0,1) \succcurlyeq (1,1)$ and, in general,

$(2m,2n) \succcurlyeq (2m+1,2n), (2m,2n+1) \succcurlyeq (2m+1,2n+1)$ for all $m,n \in \mathbb{Z}$.

So continuity at x boils down to $x \preccurlyeq y \Rightarrow f(x) \preccurlyeq f(y)$ for all $y \in \mathbb{Z}^2$; continuity everywhere to the same implication but now for all $x,y \in \mathbb{Z}^2$. 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 $x,y \in \mathbb{Z}$ 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 $f\colon \mathbb{Z} \to \mathbb{Z}$, 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 $x,y \in \mathbb{Z}^2$ 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

A continuous function $f\colon \mathbb{Z}^2 \to \mathbb{Z}$ is Lip-1 for the l^∞ norm. More generally, the conclusion holds for any continuous function $f\colon X \to \mathbb{Z}$, where X is a connected subset of $\mathbb{Z}^2$ 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 $y \preccurlyeq x$.

Theorem

A function $f\colon \mathbb{Z}^2 \to \mathbb{Z}$ is continuous if and only if it is separately continuous. More generally, the equivalence holds for any function $f\colon X \to \mathbb{Z}$ where X is a subset of $\mathbb{Z}^2$ 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.

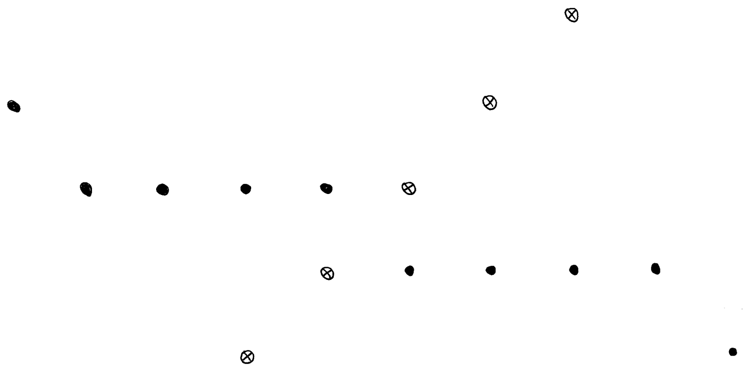
The case when $x_1 \neq y_1$ and $x_2 \neq y_2$ remains to be considered. Then $y_j = x_j \pm 1$. One of the points $q(x,y)$ and $q(y,x)$ belongs to X; let z be one of them that does. Then clearly $x \preccurlyeq z \preccurlyeq y$, which in view of the separate continuity implies $f(x) \preccurlyeq f(z)$ and $f(z) \preccurlyeq f(y)$, and we are done. □

Like in real analysis there is an intermediate-value theorem for the Khalimsky line:

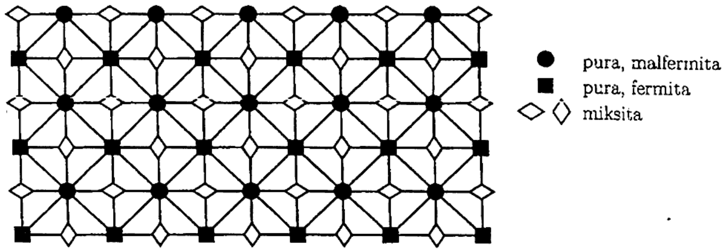
Theorem

Let two continuous functions $f, g\colon I \to \mathbb{Z}$ be given on a Khalimsky interval $I = [a, b]_{\mathbb{Z}}$. 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 $f(p) = g(p)$.

In particular this means that two Khalimsky lines which are not parallel intersect.



Two digital straight lines without a common point.



Here digitizations must move along the drawn lines.

Erik Melin's digitization respecting the Khalimsky topology.

Shiva Samieinia's unified treatment of the Khalimsky chord property and the chord property for 8-connectedness.

Thank you for your interest!

I hope we meet again!