Theoretical Computer Science 406 (2008) 146–159

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

On isoperimetrically optimal polyforms

Daniel Vainsencher*, Alfred M. Bruckstein

Computer Science Department, The Technion 32000 Haifa, Israel

ARTICLE INFO

Keywords:
Discrete geometry
Isoperimetric inequality

ABSTRACT

In the plane, the way to enclose the most area with a given perimeter and to use the shortest perimeter to enclose a given area, is always to use a circle. If we replace the plane by a regular tiling of it, and construct polyforms i.e. shapes as sets of tiles, things become more complicated. We need to redefine the area and perimeter measures, and study the consequences carefully. A spiral construction often provides, for every integer number of tiles (area), a shape that is most compact in terms of the perimeter or boundary measure; however it may not exhibit all optimal shapes. We characterize in this paper all shapes that have both shortest boundaries and maximal areas for three common planar discrete spaces.

1. Introduction

In the continuous plane $\mathbb{R}^2$, disks are special in that they have maximal area for a given length of boundary. They also have minimal length of boundary, for any given area. These two claims are equivalent in $\mathbb{R}^2$ because disks can be scaled arbitrarily to have any boundary length or any area. It is precisely this fact that also makes disks the only shapes that are optimal in those senses. These facts may be summarized by a theorem called the isoperimetric inequality:

Theorem 1. *Let γ be a simple rectifiable closed curve in $\mathbb{R}^2$, then by the Jordan theorem, it encloses a finite area A. If we denote the length of the curve L, then $A \leq \frac{L^2}{4\pi}$.*

Digital geometry is concerned, among other things, with regular tilings of the plane, for example by squares or hexagons and the shapes that arise by considering sets of tiles, also called polyforms. For example, digital straight lines (DSLs) [4] may be defined as subsets of $\mathbb{Z}^2$ that fulfill a linear inequality: $\{(i, j) \in \mathbb{Z}^2 | j \leq ai + b\}$, and their interesting properties arise from the interaction between the reals a, b and the integer grid.

In digital geometry, we can scale any shape up by any integer number n, by replacing each tile with a "meta" tile of "side size" n, and this results in a valid shape. However, this kind of scaling is usually impossible for non-integer ratios. Even worse, this form of scaling does not generally preserve geometric properties that we may care about. For example, scaling the digital straight line $\{(i, j) | j \leq i\}$ by a factor of 2 on $\mathbb{Z}^2$ does not result in a digital straight line.

Therefore, it should not come as a surprise that the above discussed equivalence between the two interpretations of the isoperimetric inequality in $\mathbb{R}^2$ does not extend to discrete cases. The discrete case reveals its secrets via a delicate analysis of the interaction between suitably defined concepts of boundary size (in two dimensions, this is the perimeter) and area for each regular tiling of the plane. Before we present one existing approach to such problems and describe our own results, we shall define our aims and terms more precisely on a particular example.

We identify the tiling of the plane by squares of unit length sides with $\mathbb{Z}^2$. It is then natural to define the area of a shape as the number of tiles in it. When it comes to defining its perimeter, however, there is more than one natural way

* Corresponding address: Computer Science Department, The Technion, Technion City, 32000 Haifa, Israel. Tel.: +972 544633624.
E-mail address: danielv@cs.technion.ac.il (D. Vainsencher).

0304-3975/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2008.06.043

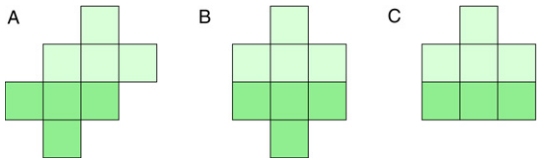

Fig. 1. Three shapes with optimal boundary for their size. *A* and *B* show non-uniqueness of optimal shapes, *C* can be enlarged without affecting the boundary size.

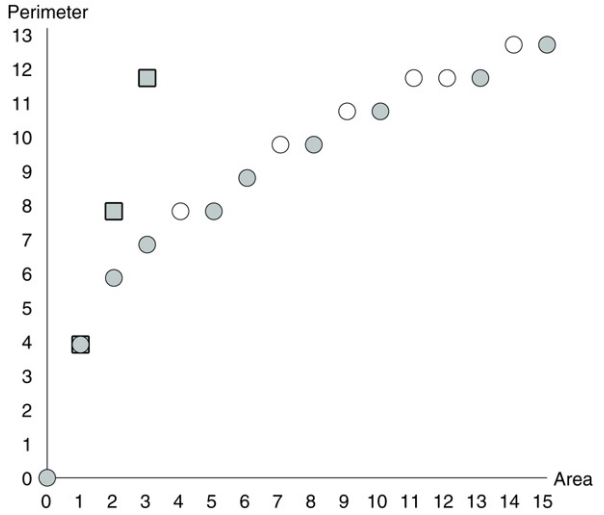

Fig. 2. Feasible $(a(S), p(S))$ pairs with maximal and minimal boundaries, delimiting F in X_4. The squares have maximal perimeters, the circles have minimal perimeters, the full circles also have maximal area for their perimeter. For example a single tile has 4 neighbors and is extremal in both ways. Thus every shape S with area $a(S)$ must have perimeter $p(S)$ between a circle and a square.

to do so. Considering length of boundary literally, we might count the number of sides that squares inside the shape share with squares outside it. Another natural definition of boundary size is to count the number of squares that are neighbors of the shape, where neighbors may be defined as sharing a side, or as sharing a side or corner with squares in the shape. Unfortunately these definitions are not equivalent, so we must select one of them before we can explore the relation between the area of a shape, and the size of its boundary.

For example, let us choose to define the boundary size as the number of squares sharing a side with the shape. Now we can consider the set of geometrically distinct shapes in Fig. 1, and note the following: *A* and *B* are obviously different shapes with the same area and the same neighborhood size.

Two questions naturally arise.

(1) Is there a unique shape of maximal area for a given boundary size, or a unique shape of minimal boundary size for a given area? The answer is no, because as will become clear later, *A* and *B* are optimal in both senses.
(2) Is every shape of maximal boundary for its area also of maximal area for its boundary size? Again the answer is no, because it is easy to add a square to shape *C* without increasing its boundary size, despite the fact (which is harder to see) that *C* has optimal boundary size for its area.

Therefore, given the above definitions, the two optimality conditions are clearly not equivalent. The relations between shapes, areas and (at least) this definition of boundary size are more complicated than those in the continuous case.

From a general perspective, we have a set O of objects of interest, i.e. the shapes or subsets of a tiling, and two ways to measure the size of an object, i.e. the area and boundary size (or perimeter). Hence, to each shape S we associate its area $a(S) \in \{0\} \cup \mathbb{N}$ and its perimeter $p(S) \in \{0\} \cup \mathbb{N}$. Then, to learn about the interactions between these two measures for a given tiling and suitable definitions of $a(S)$, $p(S)$, we may look at the set of feasible points $F = \{(a(S), p(S)) \,|S \in O\} \subset (\{0\} \cup \mathbb{N})^2$.

For example, in the space we considered above, each square of a shape has at most 4 neighbors, therefore $p(S) \leq 4 \cdot a(S)$, equality being achieved by shapes in which every two tiles are far enough to not share neighbors. This shows a fundamental difference between this space and $\mathbb{R}^2$, in which there exist shapes with finite area but infinite boundary length, such as the Koch snowflake or other fractals [7].

Let us return to the isoperimetric issues. Since the shapes in a tiling also correspond to shapes in $\mathbb{R}^2$, we might reasonably expect that there exists a constant γ such that $\sqrt{a(S)} \cdot \gamma \leq p(S)$, for reasonable definitions of perimeter. Then a diagram of F is presented in Fig. 2. In this diagram, the lower boundary, representing the price in perimeter to be paid for increasing the area, raises various questions of interest. The set of feasible pairs $(a(S), p(S))$ along this boundary corresponds to extremal

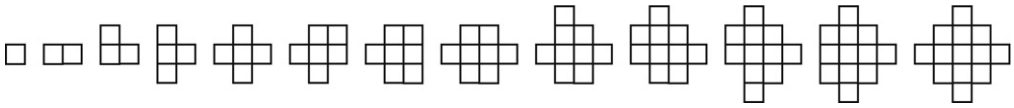

Fig. 3. A prefix of the Wang and Wang spiral.

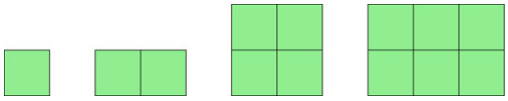

Fig. 4. Shapes generating all optimal shapes for squares, when neighbors share a side or corner.

shapes highlighting the discrete isoperimetric inequality. Note that it is possible that a feasible pair is satisfied by more than one shape, then we shall be interested in mathematically characterizing those shapes. We shall deal in this paper with the set of shapes that, like disks in $\mathbb{R}^2$, are optimal both in having the maximal area for their perimeter and in having the minimal perimeter for their area. These shapes are said to form the so-called Pareto optimal frontier of the diagram of F in Fig. 2. Similar diagrams and the concept of Pareto Optimality [8] are indeed used in economics and in engineering in situations where conflicting criteria need to be jointly optimized.

A common way of generating shapes on the boundary of F defining the isoperimetric inequality corresponding to a particular tiling, is a spiral construction. It starts with one tile, and sequentially adds tiles in a circular order growing the shape layer by layer. Harary and Harborth [6] use spirals to explore the extremal "animals" (simply connected subsets of tilings of the plane by regular triangles, hexagons and squares), which minimize the length of the perimeter for a number of tiles. Wang and Wang [14] provided a "spiral like" construction for the space $\mathbb{Z}^n$ for any n, generating an infinite sequence $T^{(n)} = (x_i)_{i=1}^{\infty}$ of tiles $x_i \in \mathbb{Z}^n$, such that any finite prefix of this sequence corresponds to a shape that has maximal compactness in the sense of having minimal boundary size for its number of $\mathbb{Z}^n$ tiles. The initial part of this construction for $\mathbb{Z}^2$ is given in Fig. 3.

As discussed above, "spiral constructions" are known to yield isoperimetric inequalities and thus the lower boundary of the F diagram for the corresponding discrete spaces. Although very pleasing, these constructions leave several interesting geometric questions open, and we shall address some of them herein.

Let us look at the set of shapes in Fig. 3 again. Since $p(S)$ (the perimeter) increases more slowly than $a(S)$, and both are in $\mathbb{Z}_+$, not every shape in the sequence can have a larger perimeter than its predecessor. Therefore, not all shapes in this sequence can have maximal area for their perimeter. Which of those shapes are optimal in this sense as well? Does this sequence of shapes exhibit all shapes with smallest perimeter for their area? The answer to this second question is no. We have already seen two shapes A and B that will turn out to be Pareto optimal and have the same area, and clearly only one of them can appear in Wang and Wang's sequence, or any similar spiral construction.

2. Brief overview of results

In this work we shall completely characterize the set of Pareto optimal shapes for three planar discrete spaces. In the case considered above ($\mathbb{Z}^2$ with the perimeter defined as the number of neighboring tiles sharing a side with the shape) our characterization will also exhibit Pareto optimal shapes that are not generated by the spiral construction. In the other two cases we shall prove the opposite: that a spiral construction does generate all the Pareto optimal shapes.

Definition 2. The expansion of a shape is the union of the shape with its neighbors.

Expansions can obviously be iterated to create sequences of shapes.

In the tiling of $\mathbb{R}^2$ by squares, we can define the perimeter of a shape as the number of tiles outside of it that share with it **a side or a corner**. This we call the eight neighbor boundary, and its association with the tiling we call it the eight-connected grid, or sometimes X_8.

Theorem 3. *A shape in the eight-connected grid is Pareto optimal if and only if it is the empty shape, or one of the shapes in Fig. 4 or generated by iterated expansions of such a shape.*

In the tiling of $\mathbb{R}^2$ by squares, we can define the perimeter of a shape as the number of tiles outside of it that share with it **a side**. This we call the four neighbor boundary, and its association with the tiling we call it the four-connected grid, or sometimes X_4.

Theorem 4. *A shape in X_4 is Pareto optimal if and only if it is the empty shape, or one of those shown in Fig. 5, or it is generated by iterated expansions of such a shape.*

Theorem 5. *Consider the tiling of $\mathbb{R}^2$ by hexagons, defining the perimeter of a shape as the number of tiles outside of it that share with it **a side**. Then a shape is Pareto optimal if and only if it is the empty shape, or one of those shown in Fig. 6, or it is generated by iterated expansions of such a shape.*

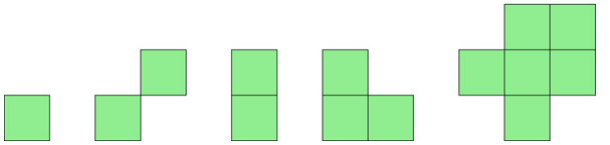

Fig. 5. Shapes generating all optimal shapes for squares, when neighbors share a side.

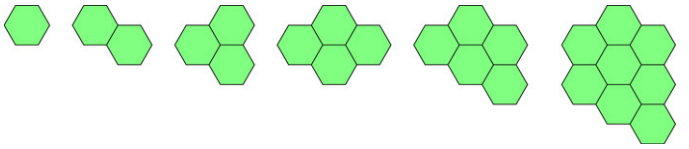

Fig. 6. Shapes generating all optimal shapes for hexagons.

Brunvoll et al. [5] mention the same shapes, in the interesting context of organic chemistry. They study planar molecules with formulas C_nH_s, which consist of hexagonal rings of carbon, whose otherwise free bonds are taken by the hydrogen atoms. The numbers n and s of carbon and hydrogen atoms respective are determined by the number of hexagons in the shape and the length of its perimeter. Interestingly, when the area is maximized with respect to this clearly different boundary measure, the same optimal shapes are obtained.

We recall that digital straight lines were defined as half planes governed by real-valued coefficients. It is not hard to see that each of the initial shapes above is generated by the intersection of several DSLs. We shall see below in the proofs of these results that all the optimal shapes can be described so, and the coefficients of the DSLs are very particular ones.

Another way to describe the Pareto frontier is to consider the sequences of $(a(S), p(S))$ pairs of Pareto optimal shapes. The sequences for neighborhood sizes are boring in that they include almost all natural numbers — after a short finite prefix, the 8 neighborhood case allows exactly all even perimeters, and the others allow all perimeters. As we might expect, the sequences for area grow more or less quadratically in their perimeter for all of these spaces, and there are some interesting connections.

Theorem 6. *For the square tiling with the 8-neighbor boundary, for any $n \in \mathbb{N}$, the area of a non-empty optimal shape with the nth smallest feasible perimeter is the maximal product of two integers whose sum is n.*

The characterization as the maximal product of two integers whose sum is n is given by Sloane in his wonderful Encyclopedia of Integer Sequences [11] for the sequence identified there as A002620. The first few values of this sequence are as follows: 0, 0, 1, 2, 4, 6, 9, 12, 16, 20, 25, 30, 36, 42, 49, 56, 64, 72. Note that the theorem above uses $n \geq 1$, so 0 appears only once.

Theorem 7. *For the square tiling with the 4-neighbor boundary, the sequence of areas of non-empty optimal shapes in order of increasing boundary size is given by $B(0,m) \leq B(1,m) \leq B(2,m) \leq B(3,m) \leq B(0,m+1) \leq B(1,m+1) \leq \cdots$ for $m \in \{-1,0\} \cup \mathbb{N}$, where $B(i,m)$ is defined as follows:*

For $i = 0$, $B(0,m) = 2m^2 + 6m + 5$.

For $i \in \{1,2,3\}$, $B(i,m) = 2m^2 + (6+i)m + 4 + 2i$.

This sequence of areas begins with 1, 2, 3, 5, 6, 8, 10, 13, 15, 18, 21, as seen in Fig. 2. Consider the sequence $\left\lfloor \frac{n^2}{8} + \frac{1}{2} \right\rfloor$, which begins 0, 0, 1, 1, 2, 3, 5, 6, 8, 10 ... and is numbered A001971 in the Sloane encyclopedia. The agreement between these sequences (starting from the fourth element of the latter) is not limited to the finite prefix we show here, it is described and proved in in Section 4.3. A better, geometric proof of this fact remains a challenge for the future.

Theorem 8. *For the hexagonal tiling, the sequence of areas of non-empty optimal shapes in order of increasing boundary size is given by one occurrence of $A(1,0)$ followed by $A(2,j) \leq A(3,j) \leq A(4,j) \leq A(5,j) \leq A(1,j+1) \leq A(6,j) \leq A(2,j+1) \leq \cdots$ for $j \in \{0\} \cup \mathbb{N}$, where $A(i,j)$ is defined so:*

For $i = 1$, $A(i,j) = 3j^2 + 3j + 1$.

For $i \in \{2,3,4,5\}$, $A(i,j) = 3j^2 + (3+i)j + i$.

For $i = 6$, $A(i,j) = 3j^2 + 10j + 8$.

This sequence begins with 1, 2, 3, 4, 5, 7, 8, 10, 12, 14, 16, 19, 21, 24, 27, 30 and is already known as sequence A001399. One of the characterizations given for it is: the numbers of the tiles along a spiral at which "folds" occur, which is equivalent to the areas of the shapes occurring during that construction that have no extra corners, so that they can be defined naturally by 6 DSLs.

Let us next proceed to prove the above results. We shall analyze the eight-connected grid first, and it will serve as a template for the more complex spaces. Some details omitted for lack of space may be found in a technical report [13].

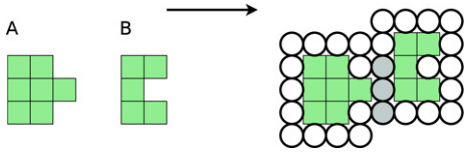

Fig. 7. Any two shapes can be translated to share 3 neighbors in X_8.

3. Pareto optimality on the eight-connected grid

3.1. Optimality conditions

We consider $\mathbb{Z}^2 = \{(i,j)\,|i,j \in \mathbb{Z}\}$ as the vertices of a graph, which has an edge between (i,j) and (m,n) iff $\max\{|i-m|, |j-n|\} = 1$. Thus every vertex has 8 neighbors. We define $\mathbf{d}_8(x,y)$ to be the distance between $x, y \in \mathbb{Z}^2$ induced by this graph, which is simply the length of the minimal path between x and y, and equal to $\max\{|i-m|, |j-n|\}$. We note that $\mathbf{d}_8$ is a natural measure of distance since it fulfills the requirements of a metric, including the triangle inequality. We will analyze shapes in the space consisting of the points of $\mathbb{Z}^2$ with the distance $\mathbf{d}_8$, denoted $X_8 = \left(\mathbb{Z}^2, \mathbf{d}_8\right)$, and called the eight-connected grid. We will construct other spaces by varying the set of points and the distance function. In the following discussion we assume $X = X_8$, though we will later see that many of the definitions and claims can be generalized to other spaces.

Definition 9. A shape A in X is a finite subset of X. The neighborhood or boundary of A is $N(A) = \{x \in X | d(x,A) = 1\}$.

Definition 10. We call a shape A area optimal if it has the maximal area for its boundary size, or formally, if $\forall B\,((|N(A)| \geq |N(B)|) \Rightarrow (|A| \geq |B|))$.

Definition 11. We call a shape A boundary optimal if it has a minimal boundary area for its area, or formally, if $\forall B\,((|A| \leq |B|) \Rightarrow (|N(A)| \leq |N(B)|))$.

We shall call a shape optimal if it is Pareto optimal, meaning that it is both area optimal and boundary optimal. Our goal will be to characterize the set of optimal shapes.

The isometries of X are those mappings $X \to X$ that preserve distances so f is an isometry if $(\forall x, y \in X)\, d(x,y) = d(f(x), f(y))$. Note that isometries of X are bijections, and also preserve the sizes of neighborhoods, therefore they preserve also optimality. The isometries of X_8 include all finite compositions of the mappings $(x,y) \mapsto (-x,y)$ (reflection), $(x,y) \mapsto (y,-x)$ (rotation by $\frac{\pi}{2}$) and $(x,y) \mapsto (x+1,y)$ (translation).

Note that these definitions are general, and will make sense even if we replaced $\mathbf{d}_8$ by a different metric, induced by a different neighbor relation, and possibly with a different set of isometries. We shall indeed analyze two other spaces using many of the same definitions and techniques, after we have found the optimal shapes of X_8.

3.2. Some suboptimal shapes and notation

Given any shape A, we can find the rightmost column on which it has elements, and denote $a_{r,b} = (x,y)$ the element of that column which is lowest (in A's the rightmost column, take the bottom element). Given a shape B, we can similarly find $b_{l,b}$ (in the leftmost column, the bottom element). Then there exists a translation T, such that $T\left(b_{l,b}\right) = (x+2, y)$. By the construction of T and the definitions of $a_{r,b}$ and $b_{l,b}$, we find that $d(A, T(B)) = 2$, and that $|N(A) \cap N(T(B))| \geq \left|N\left(a_{r,b}\right) \cap N\left(b_{l,b}\right)\right| = 3$. An example of this construction is given in Fig. 7.

We state a slightly weaker conclusion formally:

Lemma 12. *For any two shapes A, B in X, there exists a translation T such that $A \cap T(B) = A \cap N(T(B)) = N(A) \cap T(B) = \emptyset$ so that $|A \cup T(B)| = |A| + |T(B)| = |A| + |B|$, and also $|N(A \cup T(B))| \leq |N(A)| + |N(B)| - 2$.*

The reasoning given above actually supports $|N(A \cup T(B))| \leq |N(A)| + |N(B)| - 3$, but the lemma as given holds across other spaces we shall discuss. We shall use this lemma to diagnose non-optimal shapes.

Lemma 13. *Let A be the union of non-empty sets B, C such that $B \cap C = B \cap N(C) = N(B) \cap C = \emptyset$ (so each part overlaps neither the other part nor its neighbors), and $|N(B) \cap N(C)| \leq 1$. Then A is not optimal.*

This follows from the definitions, Lemma 12, and some application of the inclusion exclusion principle: $|A \cup B| = |A| + |B| - |A \cap B|$.

We now take some arbitrary $x_{\text{orig}} \in X$ as our origin. x_{orig} has eight neighbors, we name them $x_0, \ldots x_7$, moving counter clockwise from the neighbor to the left and below x. We represent directions in this space as $d_i = x_i - x_{\text{orig}}$ (Fig. 8), and say that i is the direction index of d_i. Denoting $\mathbb{A} = \{1,3,5,7\}$, $\{d_i\}_{i \in \mathbb{A}}$ correspond to directions parallel to the axes. It is natural to treat the direction indices cyclically, so that the direction following d_7 is d_0. One way to state this is to consider direction indices as belonging to the finite group $\mathbb{Z}/8\mathbb{Z}$ in which $7 + 1 = 0$. Then $d_{i+4} = -d_i$ is the direction exactly opposed to d_i.

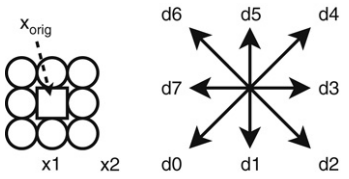

Fig. 8. The neighbor directions in X_8, induced by an arbitrary tile x_{orig} and its neighbors x_1, x_2 and so forth.

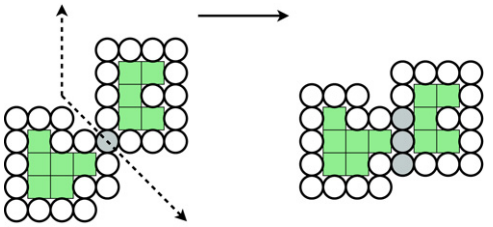

Fig. 9. A union of separated shapes is not optimal.

Definition 14. Let i be a neighbor direction index (so that $d(x, x + d_i) = 1$), then the ray from x_0 in direction i is the set $R_{x_0,i} = \{x_0 + pd_i | p \in \{0\} \cup \mathbb{N}\}$.

Remark 15. Clearly, $R_{x_0,i}$ is a connected subgraph of X. Note that connectedness in this paper always refers to the graph theoretic concept, not a topological one.

Definition 16. Let $i \neq j$. We define the cut along $R_{x_0,i}$ and $R_{x_0,j}$ to be the set $L_{x_0,i,j} = \{x_0\} \cup R_{x_0,i} \cup R_{x_0,j}$.

Again, $L_{x_0,i,j}$ is always a connected shape. This does not, however, guarantee that $X \backslash L_{x_0,i,j}$ is not also connected, for the same reason that the two diagonals of a chess board are each connected in this way, despite crossing. This is a peculiarity of the $\mathbb{Z}^2$ grid which can be solved by considering the complementary metric for background images [12]. For our purposes it is enough to note that a cut $L_{x_0,i,j}$ in X_8 partitions $X \backslash L_{x_0,i,j}$ into two infinite sets that do not share a side.

Definition 17. Let $L_{x_0,i,j}$ be a cut in X_8. We can write $X \backslash L_{x_0,i,j} = A \cup B$ where A, B are disjoint, infinite, and do not share a side. Then we shall say that A, B are separated by $L_{x_0,i,j}$.

Lemma 18 (*The Separation Lemma*). *Let A be a shape, and let B, C be the two parts of $X \backslash L_{x_0,i,j}$ so that $N(A) \cap B$ and $N(A) \cap C$ are not empty, and $\left|N(A) \cap L_{x_0,i,j}\right| \leq 1$. Then A is not optimal.*

This formalizes a fact that seems almost obvious from the inspection of Fig. 9.

Proof. First we will show that $A \cap L_{x_0,i,j} = \emptyset$, so that A is the disjoint union of subsets of B, C. Assume that $a \in A \cap L_{x_0,i,j}$, then a has at least two neighbors along $L_{x_0,i,j}$ (because i, j are neighborhood directions). Each of them either is, or is not in A. If it is not, then we have found an element of $N(A) \cap L_{x_0,i,j}$. If it is in A the same argument can be applied again. This will only occur a finite number of times because A is finite, then eventually we will find a neighbor of A in each of the two directions along $L_{x_0,i,j}$. So the presence of a entails $\left|N(A) \cap L_{x_0,i,j}\right| \geq 2$, which contradicts a given assumption.

Let $n \in X \backslash L_{x_0,i,j}$, wlog we assume that $n \in B$. Now let us assume also that $n \in N(A \cap C)$, and show that this leads to a contradiction. We have seen that in X_8, elements of B and of C cannot share a side (then in particular, this is true of n and each element of $A \cap C$). Then because $n \in N(A \cap C)$, they must share a corner, we assume wlog that a_C is the tile in $A \cap C$ sharing this corner with n. Then n, a_C are on a diagonal crossing $L_{x_0,i,j}$. Then a_C shares sides with two elements of $L_{x_0,i,j}$, in contradiction to given assumptions. This shows that $A \cap C$ does not have a neighbor across $L_{x_0,i,j}$. Clearly the same holds for $A \cap B$.

From this we conclude three facts:

(1) Any element of $N(A) \cap B$ is a neighbor of $A \cap B$, and similarly any element of $N(A) \cap C$ is a neighbor of $A \cap C$. In particular, neither of $A \cap C$ and $A \cap B$ is empty.
(2) Any neighbors shared by $A \cap B$ and $A \cap C$ must be in $L_{x_0,i,j}$.
(3) Neither of $A \cap B, A \cap C$ can overlap a neighbor of the other.

Then we have shown that A is a disjoint union of subsets of B, C, neither subset is empty, each of $A \cap B, A \cap C$ does not overlap the others neighborhood, and the only possible common neighbor of both is in $L_{x_0,i,j}$. Then we can apply Lemma 13 to conclude that A is not optimal.

Remark 19. Note that most of this proof depended only on Lemma 13 and definitions. The only point at which we considered features particular to X_8 was when noting that a subset of A on one side cannot have a neighbor from the other.

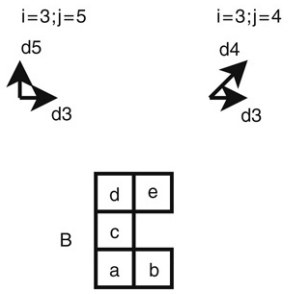

Fig. 10. A shape and two coordinate systems.

Lemma 18 will turn out to be very easily usable in proving that optimal shapes have specific forms. Here is a sketch of its use. Assume that we have found $L_{x_0,i,j}$ such that x_0 is the only neighbor of a shape A along $L_{x_0,i,j}$. Assume also that each of the connected components of $X \backslash L_{x_0,i,j}$ contains an element of $N(A)$, then by the lemma A is not optimal.

At the beginning of this section, we expended some effort to specify particular elements in shapes. In the rest of this paper, we will find ourselves doing this more than once, so it is useful to take a moment to introduce some concise and general notation.

Let $i \in \mathbb{A}$ then d_i (called the main direction) is parallel to the axes. Let $j \notin \{i, i+4\}$ (so d_j is neither d_i nor its inverse). Any such choice induces a coordinate system on X, because for each x there exists unique $p, q \in \mathbb{Z}$ such that $x = x_{\text{orig}} + pd_i + qd_j$. We cannot have both directions be diagonal, because then coloring the plane as a chess board, we would find that we can represent using integer coefficients only those tiles sharing the color of the origin.

Fig. 10 shows two examples of coordinate systems induced by such choices of directions. One is rather standard, and the other is not. In the second, if we set $a = x_{\text{orig}}$, then $b = x_{\text{orig}} + 1d_i + 0d_j$, $c = x_{\text{orig}} - 1d_i + 1d_j$, $d = x_{\text{orig}} - 2d_i + 2d_j$, $e = x_{\text{orig}} - 1d_i + 2d_j$.

Definition 20. Given i, j as described above, we define the function $\phi_{i,j} : X \to \mathbb{Z}^2$ as $x \mapsto (p, q)$ where $x = x_{\text{orig}} + pd_i + qd_j$. The function $\phi_{i,j}$ is called the coordinate mapping with directions i, j.

So $\phi_{i,j}$ assigns to every element in X its coordinates in the coordinate system defined by x_{orig}, i, j. We will always assume x_{orig} to be constant throughout, and only i, j vary, so $\phi_{i,j}$ is well identified.

Note that in our example, for $j = 4$, tiles c, e have the same first coordinate. So the second direction j, which can be diagonal or not, allows us to divide the plane into equivalence classes, each of which is a line in the direction d_j. The direction i induces an order on these lines.

Definition 21. Assume a coordinate mapping with the directions i, j and a shape A. Partition A into layers $A_k = \left\{x \in A | \phi_{i,j}(x) \in \{k\} \times \mathbb{Z}\right\}$ so that all elements of a layer have the same coordinate in the direction i, and vary in the j coordinate. Let $k_{\max} = \max_{A_k \neq \emptyset} k$ so that $A_{k_{\max}}$ is the layer with maximal k that is not empty. Then $\psi_{i,j} : \mathcal{P}(X) \to X$ is defined so that $\psi_{i,j}(A)$ is the element of $A_{k_{\max}}$ with maximal j coordinate.

In our example, $\psi_{3,5}(B) = e$ and $\psi_{3,4}(B) = b$. Using this notation to describe the beginning of this section, we could have written simply that $a_{r,b} = \psi_{3,1}(A)$ and $b_{l,b} = \psi_{7,1}(B)$.

3.3. The optimal shapes on the eight-connected grid are rectangles

Starting from the next subsection, we will consider the optimality of shapes from a limited set of rather simple shapes.

Definition 22. We call A a rectangle if $A = \{(a, b) | a \in [c, d] \wedge b \in [e, f]\}$. In this case, we say that the dimensions of A are $j = d - c + 1, k = f - e + 1$.

We note that the dimensions of a rectangle define it up to translation. Optimality does not depend on location, but only on dimensions, so we will consider only those. Similarly optimality is invariant to rotation by a quarter turn, so we can assume that $j \leq k$. Our task of characterizing the optimal shapes will clearly be simpler when we restrict ourselves to the rectangles, but to justify doing so, we must first prove that they are sufficient.

Lemma 23. *Any optimal shape in X_8 is a rectangle.*

We shall perform this step of restricting the set of optimal shapes to simple shapes defined as the intersection of half planes in each of the spaces we consider, and in each space the proof uses the same geometric ideas, though the details of the construction differ. Before considering in detail X_8, here are the considerations we will apply (Figs. 11 and 12 illustrate them):

(1) We assume that there exists an optimal non-simple shape A. To show this leads to a contradiction, we consider the minimal simple shape B s.t. $A \subset B$. Then $A \neq B \Rightarrow |A| < |B|$.

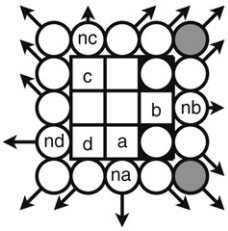

Fig. 11. In this example A is as before, B is its bounding rectangle (black background). $a = a_1$; $b = a_2$; ... similarly $na = n_1 \ldots$ The white circles are in $N(A)$, the grey ones are in $N(B)\backslash N(A)$. Here $N_{5,7}$ consists of the two circles between nd and na. The direction associated with each element of $N(B)$ is shown as an arrow.

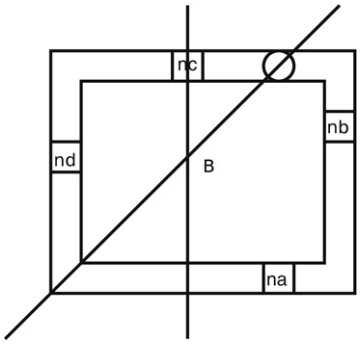

Fig. 12. Diagram of a separation lemma argument, if $p(n)$ is not finite. Each line divides the plane into a left half, which we call C, and a right half called D. The vertical line is created when $n = nc = n_5$, then each of C, D contains one of nb, $nd = n_3, n_7$. The diagonal line corresponds to the case in which $n \in N(B)$ is not one of $\{n_i\}$, but in $N_{3,5}$ placing each of the delimiters of $N_{3,5}$, which are nb, $nc = n_3$, n_5 in one of C, D.

(2) Then if we show a mapping $m : N(B) \to N(A)$ that is one to one, then B has no more neighbors than A does, then A is non-optimal.
(3) We construct this mapping by assigning to each element n in $N(B)$ a direction into B. We will later choose the direction of each ray more precisely, so that it points towards A. We then define its end point $m(n)$ to be the first element of $N(A)$ encountered by a ray from n along this direction.
(4) In order to make the mapping be well defined and one to one, use the following geometric facts to choose the assigned directions:
   (a) Rays from adjacent elements of $N(B)$ which are in parallel directions cannot meet. Note however, that this alone cannot suffice because we must change directions somewhere.
   (b) B is minimal, therefore some extremal elements of A are also extremal in B. So B shares some neighbors with A. If $n \in N(A) \cap N(B)$, then a ray from n also ends in n, never leave it and is in no danger of collision with other rays. This allows us to change the assigned directions wherever A is adjacent to $N(B)$.
   (c) For a moment suppose that a ray divides $N(B)$ into two parts, but never meets an element from $N(A)$, then our mapping might not be well defined, because the ray does not end. But our ray is into B, so it intersects $N(B)$ at another point (call it q) dividing it into two parts. Because the direction associated with q must also be into B, it cannot be the same as that associated with p, therefore somewhere along each of the two parts of $N(B)$, which are paths from p to q, the direction must have changed. Direction changes are only associated with points at which $N(A)$ and $N(B)$ coincide, then each component of $X\backslash L$ contains an element of $N(A)$, therefore an element of A, and we can apply the separation lemma to contradict the assumed optimality of A.
   (d) If two rays meet from different directions and their points $n_a, n_b \in N(B)$ are not adjacent, then as we argued above, on each of the two paths between them along $N(B)$, the directions changed, allowing the use of the separation lemma.
(5) We conclude that if A were optimal, then the mapping m would be well defined and one to one, proving that A is dominated by the minimal simple shape containing it B. Then optimal, non-simple shapes do not exist.

Now we will follow this plan for X_8.

Proof. Let A be some optimal shape, then let B be the minimal rectangle such that $A \subset B$. We will show a mapping $m : N(B) \to N(A)$, and then prove that it is one to one.

Let $i \in \mathbb{A}$, then d_i is parallel to one of the axes (for example, it may be the negative vertical direction), then we denote $a_i = \psi_{i,i+2}(A)$ (in this case, the rightmost element in the bottom row of A). We denote also $n_i = a_i + d_i$, then note that $n_i \in N(A) \cap N(B)$. $N(B)$ has its sides along the rectangle B, and $\{n_i\}_{i\ \mathrm{odd}}$ are each on one of these sides, and not on its corners. Thus $\{n_i\}_{i\ \mathrm{odd}}$ partition $N(B)$ into four connected components, each of which is denoted $N_{i,i+2}$ if it is bounded by n_i, n_{i+2}. An example of this construction is provided in Fig. 11. $N_{i,i+2}$ is non-empty because it contains at least a corner of

$N(B)$, so that $\{n_i\} \cup N_{i,i+2} \cup \{n_{i+2}\}$ is shaped like an "L" (as in Fig. 12). To each $n \in N(B)$ we associate a direction $d(n)$ so: if $n = n_i$, then $d(n) = d_i$, if $n \in N_{i,i+2}$, then $d(n) = d_{i+1}$.

Note that for every $n \in N(B)$, $n - d(n)$ is a member of B that n neighbors. Our mapping m will simply go further along a ray in the same direction until the first neighbor of A is reached. Formally we will define $p(n) = \min \arg_{q \in \{0\} \cup \mathbb{N}} \{n - qd(n) \in N(A)\}$, allowing us to define the mapping $m : N(B) \to N(A)$ as $m(n) = n - p(n) \cdot d(n)$. One might reasonably worry about the soundness of these definitions if there exists no $q \in \{0\} \cup \mathbb{N}$ such that $n - qd(n) \in N(A)$. We will assume that this is the case, show this leads to a contradiction, and conclude such a q must exist.

Now we consider $L = L_{n,d(n),-d(n)}$, and note that it separates X into two connected components C, D (as in Fig. 12) so that their disjoint union is $X \setminus L$, and so that $N(A) \subset C \cup D$. In particular, $\left|N(A) \cap L_{x_0}\right| = 0 \leq 1$, then if we show that each of C, D has an element of $N(A)$, we can conclude that A is non-optimal by the separation lemma, in contradiction to our assumption.

If $n = n_i$ for some odd i, then L separates the rectangle $N(B)$ into two components each containing a whole side, and in particular one of $n_{i\pm2}$. If $n \notin \{n_i\}_{i \text{ odd}}$ then we write $d_i = d(n)$, we then note that L separates $N_{i,i+2}$ into two components such that each is continued by one of n_i, n_{i+2}. In either case, each of B, C contains at least one element from $\{n_i\}_{i \text{ odd}}$.

Then the definitions of $p(n)$ and $m(n)$ are sound.

Now it remains to show that $m(n)$ is one to one. To do this, we assume that there exist distinct elements $n, n' \in N(B)$ that have $m = m(n) = m\left(n'\right)$, and show that this leads to a contradiction. We note that by the definition of m, the only element in $L_{m,d(n),d\left(n'\right)}$, that can belong to either A or $N(A)$ is $m \in N(A)$. Again we will use the separation lemma to contradict the optimality of A, and to do so we need to show each component contains an element from $N(A)$. We note that $d(n) \neq d\left(n'\right)$, because otherwise their lines are parallel and they do not have a common m.

If we can write $n \in N_{i,i+2}$ and $n' \in N_{j,j+2}$ where $i \neq j$ are odd, then n_i and n_{i+2} are separated by $L_{m,d(n),d\left(n'\right)}$.

Otherwise, at least one of the rays starts at one of the special elements n_i, and wlog we assume that $n = n_i$, then we note that $m(n_i) = n_i$, which is in $N(A) \cap N(B)$. Now we consider in turn all the possibilities for the direction j corresponding to n'. If j is $i \pm 1$, then the ray corresponding to it must arrive at n_i from outside the rectangle $B \cup N(B)$, which is inconsistent with the construction. Similarly, $j = i \pm 2$ would require $n' = n_j$ to be at a corner of $N(B)$ which is again inconsistent with the construction, because then $n_j - d_j \notin A$. For other j, L_{m,d_j,d_i} admits a separation argument, using n_{i-1} and n_{i+1}. □

3.4. The optimal rectangles on the eight-connected grid have almost equal sides

In this section, knowing that all optimal shapes are rectangles, we will pause to relate our notation to well-known facts about rectangles, then find a subset of rectangles that includes all optimal ones, and then prove that this subset is exactly the set of optimal shapes.

Lemma 24. *A rectangle of dimensions j, k, has area $|A| = j \cdot k$, and boundary size $\left|A'\right| = 2j + 2k + 4 = 2 \cdot (j + k + 2)$.*

These facts and simple calculation allow us to easily verify the following:

Proposition 25. *A rectangle of dimensions j, k is optimal iff $|j - k| < 2$.*

We already know that $|j - k| < 2$ is a necessary condition for optimality, now we want to show that it is sufficient as well. In other words, that for no shape A fulfilling the condition, there exists a shape B strictly improving on it. To do this it is sufficient to compare A only to other shapes B that also fulfill the necessary condition — but we need to explain why. Assume that A is improved by some shape A_2 that is not optimal. Then by definition, A_2 is strictly improved by some other shape, and so on. If this chain is finite, the last shape A_n must be optimal, and we are done. Then it is enough to note that infinite chains of improvement are not possible. This is because each improvement must either decrease a perimeter by at least one and leave it positive, or increase the area by at least one. So the shapes in an infinite chain of improvements must have unbounded area. This however falls into contradiction with the isoperimetric inequality for $\mathbb{R}^2$ and the perimeter of the original shape.

Note that this argument also holds for different spaces and tilings of $\mathbb{R}^2$, as long as their tiles have bounded number of sides and length of sides. Now we return to prove the proposition.

Proof. Assume wlog that $j > k$. Then any shape with $|j - k| < 2$ has one of the forms (k, k) or $(k, k + 1)$. We define the partial order $\prec$ among these pairs using the total order $(0, 0) \prec (0, 1) \prec (1, 1) \prec (1, 2) \prec \cdots \prec (k - 1, k) \prec (k, k) \prec (k, k + 1) \prec \cdots$. Now we note that the areas agree with this order: $(k - 1) \cdot k < k \cdot k < k \cdot (k + 1)$, and the neighborhoods sizes also agree with it: $2 \cdot (k + k - 1 + 2) \leq 2 \cdot (k + k + 2) \leq 2 \cdot (k + k + 1 + 2)$. Then for any two shapes A, B fulfilling the necessary conditions, $A \prec B \Rightarrow N(A) < N(B) \wedge |A| < |B|$, thus neither strictly improves on the other. □

Thus we can conclude:

Corollary 26. *The set of optimal shapes in the eight-connected grid is the set of rectangles whose height and width differ by at most 1.*

This directly results in Theorem 6. Theorem 3 also results, though it generates the rectangles using expansions. Expansions are used for consistency with the results of the following sections, where they arise quite naturally.

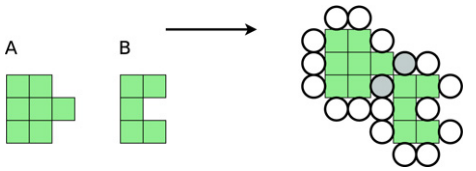

Fig. 13. Any two shapes can be translated to share 2 neighbors in X_4.

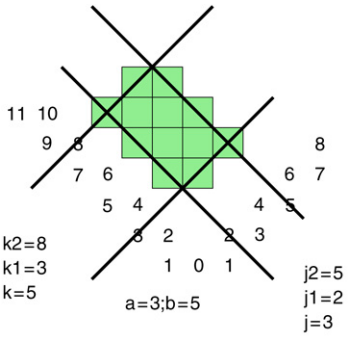

Fig. 14. A simple shape and its dimensions.

4. Pareto optimality on the four-connected grid

In this section, we will consider two squares to be adjacent only if they share an edge. Formally, let $x = (i, j)\ ; y = (m, n)$, so $x, y \in \mathbb{Z}^2$, then in our graph $G = \left(\mathbb{Z}^2, E\right)$ there will be an edge $(x, y) \in E$ iff $|i - m| + |j - n| = 1$. This induces the distance $\mathbf{d}_4\left((a, b), (c, d)\right) = |i - m| + |j - n|$, and in this section we consider $X_4 = \left(\mathbb{Z}^2, \mathbf{d}_4\right)$.

The definitions of the previous section carry over quite cleanly. Any finite set $A \subset \mathbb{Z}^2$ is a shape in X, its neighborhood is the set of squares at distance exactly one, and formally $N(A) = \left\{x \in \mathbb{Z}^2 | \mathbf{d}_4(A, x) = 1\right\}$. The new distance changes the shapes of neighborhoods, naturally affecting which shapes are optimal. As we noted, isometries must preserve neighborhoods, which again allows only translations, reflections, rotations by $\frac{\pi}{2}$ and their compositions.

The possible directions $\{d_i\}_{i=0}^{7}$ are the same as in the previous section, though now $x + d_i$ is a neighbor of x iff d_i is parallel to one of the axes, or equivalently, if $i \in \mathbb{A}$. Similarly, the definitions for $\phi_{i,j}$, $\psi_{i,j}$, rays $R_{x_0,i}$ and cuts $L_{x_0,i,j}$ are all unchanged. Note that in X_4, the neighborhood direction indices i, j allowed in rays and cuts are just those in $\mathbb{A}$. Then it is easy to see that in X_4, any cut $L_{x_0,i,j}$ separates $X \backslash L_{x_0,i,j}$ into two connected components in X_4.

In the previous section, we proved Lemma 12, used it to prove Lemma 13, which we used in turn to prove Lemma 18 for X_8. They hold also in X_4, with small modifications to two of the proofs. The first holds using the transformation $T\left(\psi_{7,0}\right) = \psi_{3,4} + d_2$, as seen in Fig. 13. The proof of the second is unchanged. As we mentioned in Remark 19, the proof of the separation lemma for X_8 applies here except for one point. Given $L_{x_0,i,j}$ such that $A \subset X \backslash L_{x_0,i,j}$, and $X \backslash L_{x_0,i,j} = B \dot{\cup} C$ (the disjoint union of B, C), and $n \in B$, we must show that $n \notin N(A \cap C)$. But the restricted neighborhood relation X_4 does not allow B, C to be neighbors at all.

In X_8, we considered rectangles which are shapes defined as the points fulfilling some inequalities on their coordinates. In X_4 we define the set of simple shapes in a similar way.

Definition 27. We call a shape A simple in X_4 if if can be written in the form $A = \{(a, b) | a + b \in [j_1, j_2] \wedge a - b \in [k_1, k_2]\}$. We will say that its dimensions are $j = j_2 - j_1$ and $k = k_2 - k_1$.

See Fig. 14. Recalling that isometries and in particular rotation by $\frac{\pi}{2}$ do not affect optimality, we may assume wlog that $j \leq k$.

Lemma 28. *Optimal shapes in X_4 are simple.*

The proof follows the outline given in the previous section, and we give here only the details that differ in the construction itself.

Let i be even, then d_i is a diagonal direction, then we denote $a_i = \psi_{i,i+2}(A)$. In this case we denote also $n_i = a_i + d_{i+1}$, then note that $n_i \in N(A) \cap N(B)$. $\{n_i\}_{i \text{ even}}$ are each on at least one of the sides of $N(B)$ (possibly more, since n_i can be in a corner). Thus $\{n_i\}_{i \text{ odd}}$ partition $N(B)$ into (at most) four connected components, each of which is denoted $N_{i,i+2}$ if it is bounded by n_i, n_{i+2}. To each $n \in N(B)$ we associate a direction $d(n)$ so: if $n \in \{n_i\} \cup N_{i,i+2}$, then $d(n) = d_{i+1}$. An example is given in Fig. 15. The definitions of p and m are not changed, and the separation lemma is used to show the one to oneness of m as before.

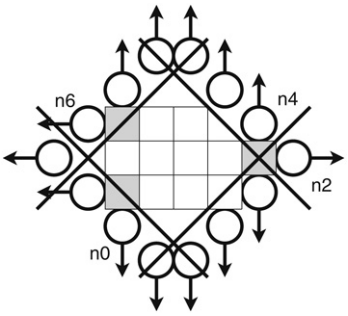

Fig. 15. A shape, its three distinct $\{a_i\}$, and the four corresponding $\{n_i\}$.

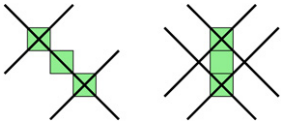

Fig. 16. A degenerate shape A, and a variation A' which clarifies that A is non-optimal.

4.1. Describing optimal simple shapes in X_4

It now remains to identify which of the simple shapes are optimal. These come in more varieties than rectangles, thus it will take a few steps, and it will be useful to introduce a few tools for the description of such shapes.

Definition 29. A simple shape A with $j = 0$ is called a *degenerate shape*.

Note that degenerate shapes behave differently from other simple shapes (for example, in the degenerate case k cannot have odd values, because the Manhattan distance between two tiles on a diagonal is always even).

Lemma 30. *The only optimal degenerate shapes have an area of* 0, 1 *or* 2.

This is clear from considering what occurs when $k = 4$: Fig. 16 shows an equivalent shape that is not simple, therefore not optimal.

We have seen that only a small and finite set of degenerate shapes is of interest to our discussion, and it has already been identified. In the following analysis of simple shapes, we consider only non-degenerate shapes.

Lemma 31. *Every simple shape has $j + k + 4$ 4-Neighbors.*

This can be seen by induction on j and on k.

Next we will describe each simple shape as a *spine* expanded by an iterative expansion process.

Lemma 32. *Let A be a simple shape of dimensions j, k. Then its expansion (as in Definition 2) is a simple shape of dimensions $j + 2$, $k + 2$.*

It is easy to see that an expansion increases the number of tiles of the shape by $j + k + 4$, and the number of neighbors by 4. See Fig. 17 for an example. This easily yields the following:

Lemma 33. *Let A be a simple shape with dimensions j, k. After s expansion steps, its neighborhood grows by 4s and its area grows by $E(j, k, s) = s(2 + j + k + 2s)$.*

Definition 34. A simple shape such that $j \in \{1, 2\}$ is called a *spine*.

Lemma 35. *A simple shape A can be described as a spine, expanded some finite number (possibly zero) of times. This description is unique.*

We shall later show that there are only 4 kinds of spines. Thus, since we know the area added by each expansion step, we will be able to calculate the areas of all simple shapes.

Lemma 36. *Let A_s be a spine of dimensions j, k, then its area is given by (See Fig. 17):*

(1) *If $j = 1$, the area is $k + 1$*
(2) *If $j = 2$, then we have the following options:*
   (a) *If k is odd, then $|A_s| = \frac{3\cdot(k+1)}{2}$.*
   (b) *k is even, of type 1, then $|A_s| = \frac{3\cdot k}{2} + 1$.*
   (c) *k is even, of type 2, then $|A_s| = \frac{3\cdot k}{2} + 2$.*

Proof. For $j = 1$, there are k tiles at distances 0 to $k - 1$ from one line, and one more. For $j = 2$, there are $\left\lfloor \frac{k+1}{2} \right\rfloor$ triplets of tiles. Note that there are two ways of getting from an odd k to an even one, depending on which boundary is moved, resulting in different area increases. □

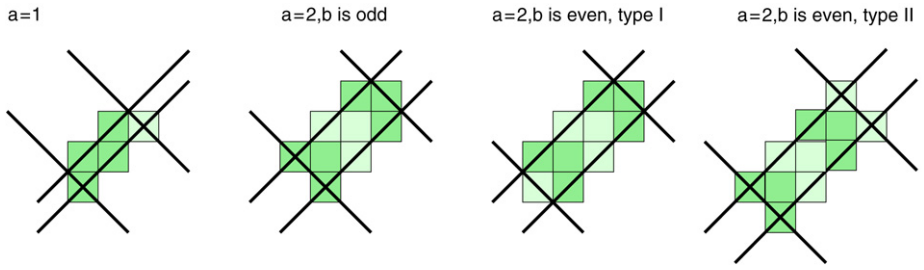

Fig. 17. Spine types and their areas.

Table 1
For each spine that may be optimal, we find i such that the perimeter of that spine after m expansions is $4(m+1)+i$

i	Spine	Spine type	Spine area	Number of expansions	Total area	Subopt.
0	(1, 3)	$a=1$	3+1	$m-1$	$2m^2+2m$	Yes
0	(2, 2)	$a=2$, b is even	$\frac{3\cdot 2}{2}+2$	$m-1$	$2m^2+2m+1$	
1	(1, 4)	$a=1$	4+1	$m-1$	$2m^2+3m$	Yes
1	(2, 3)	$a=2$, b is odd	$\frac{3\cdot(3+1)}{2}$	$m-1$	$2m^2+3m+1$	
2	(1, 1)	$a=1$	1+1	m	$2m^2+4m+2$	
2	(2, 4)	$a=2$, b is even	$\frac{3\cdot 4}{2}+2$	$m-1$	$2m^2+4m+2$	
3	(1, 2)	$a=1$	2+1	m	$2m^2+5m+3$	
3	(2, 5)	$a=2$, b is odd	$\frac{3\cdot(5+1)}{2}$	$m-1$	$2m^2+5m+2$	Yes

Then we calculate the expanded areas for those spines, and find some spines that result in shapes of smaller areas than those resulting from a different spine with the same i therefore the same perimeters.

4.2. Spines of optimal shapes are short

The facts we have found about the areas of spine types and those added by expansions allow us the following intermediate conclusion:

Lemma 37. *Let A_s be a spine with dimensions j, k of an optimal shape A, then $j+4>k$.*

Corollary 38. *The dimensions of spines of optimal shapes are a subset of:* $\{(1,1),(1,2),(1,3),(1,4),(2,2),(2,3)(2,4),(2,5)\}$

Recalling Lemma 36, we note that spines of dimensions $\{(2,2),(2,4)\}$ mentioned above come in two types. As we saw there, type 2 spines have strictly more area than those of type 1, with the same neighborhood. Therefore only type 2 spines can result in optimal shapes. In this context, each set of spine dimensions results in a certain optimal spine area and neighborhood size.

This allows us to restrict our attention to a set of shapes small enough to apply elimination.

Lemma 39. *Let A be a non-degenerate optimal shape with dimensions j, k, so that $|N(A)|=4(m+1)+i$, with $i\in\{0,1,2,3\}$. Then: $|A|=2m^2+(1+i)m+\max\{1,i\}$*

Proof. Let a, b be the dimensions of A's spine, then remembering each expansion increases the neighborhood size by 4, we see that $4(m+1)+i=j+k+4=a+b+4(s+1)$. One conclusion is that $a+b\equiv i \bmod 4$, and another is that $s=\frac{4m+i-a-b}{4}$. Hence, denoting $|A_s|$ the area of the skeleton of dimensions a, b, the total area for such a shape is exactly $|A|=|A_s|+E\left(a,b,\frac{4m+i-a-b}{4}\right)$.

Table 1 describes for each i the possible spines for optimal shapes with $|N(A)|=4m+i$, the shape's area for each spine, and the spines resulting in shapes that are suboptimal for that neighborhood size. The spines in Table 1 that are not marked as suboptimal fulfill the formula claimed. □

We have shown one necessary condition for the optimality of degenerate shapes and now another one for non-degenerate shapes. As we have done for X_8, we now introduce a partial order $\prec$ on all shapes fulfilling the necessary conditions that agrees with area and with perimeter, thus proving that the necessary conditions are also sufficient.

Naturally, we begin by ordering the 3 optimal degenerate shapes according to area. Next we note that the optimal degenerate shape of area 2 is equivalent in area and perimeter to the simple shape with spine (1, 1) and zero expansions. We continue to use the ordering of the optimal spines (see Fig. 18) by area. This ordering is extended to the other optimal shapes, which are the expansions of those given, by noting that $|A|$ is strictly monotonous in m because $2(m+1)^2+2(m+1)+1=2m^2+4m+2+2m+2+1=2m^2+6m+5>2m^2+5m+3$, and that for any m, $|A|$ and $N(A)$ are strictly monotonous in i (so that expansions preserve the order).

Lemma 40. *The non-degenerate optimal shapes are those simple shapes whose spines have one of the following forms:* $(a,b)\in\{(1,1),(1,2),(2,2),(2,3)(2,4)\}$. *These spines appear in the first row of Fig. 18.*

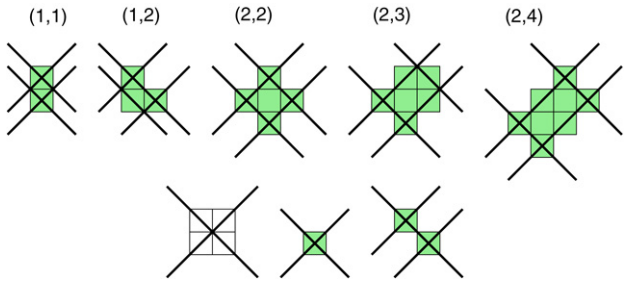

Fig. 18. The unexpanded optimal shapes.

Lemma 41. *The degenerate simple shapes with areas* 0, 1, 2 *are all optimal. These appear in the second row of Fig.* 18.

Now to prove Theorem 4, consider the union of degenerate optimal shapes and non-degenerate optimal shapes, noting that the non-empty optimal degenerate shapes, when expanded once, are equal to some of the non-degenerate optimal shapes. Theorem 7 follows by applying the formulas for area developed in Lemma 39, and the order induced by $\prec$.

4.3. Relation to the sequence of integers closest to $\frac{n^2}{8}$

Theorem 42. *Let* (A_n, p_n) *be the nth area–perimeter pair exhibited by the optimal shapes in* X_4 *(including the empty shape), when ordered with increasing area, for* $n \geq 3$. *Then* A_n *is* $\left\lfloor \frac{n^2}{8} + \frac{1}{2} \right\rfloor$.

Proof. The sequence of areas for non-empty shapes is given for $m \in \{-1, 0\} \cup \mathbb{N}$ and $i \in \{0, 1, 2, 3\}$ as follows:

(1) For $i = 0$, $B(0, m) = 2m^2 + 6m + 5$.
(2) For $i \in \{1, 2, 3\}$, $B(i, m) = 2m^2 + (6 + i)\, m + 4 + 2i$.

The proof is by induction on m for each of the four values of i. Thus the base case should consider each i. $n = 3$ requires us to consider the third feasible pair (A_3, p_3), which is $(2, 6)$. This corresponds to a spine of dimensions $(1, 1)$, and parameters $i = 2$; $m = -1$, so that $B(2, -1) = 2$ is the desired area. Then all that remains is to see that this agrees with $\frac{(n+1)^2}{8} = \frac{4^2}{8} = 2$. The other cases can be verified similarly by substitution as follows:

(1) $A_4 = 3$ has spine $(1, 2)$, and parameters $i = 3$; $m = -1$.
(2) $A_5 = 5$ has spine $(2, 2)$, and parameters $i = 0$; $m = 0$.
(3) $A_6 = 6$ has spine $(2, 3)$, and parameters $i = 1$; $m = 0$.

Induction step. We assume that the claim holds for i, m, and prove it for i, $m + 1$. In particular, if i, m corresponds to A_n, then i, $m + 1$ corresponds to A_{n+4}, and so should the areas. Note that i, m corresponds to n iff $n = 1 + i + 4\,(m + 1)$.

Now we note that $\frac{n^2}{8} - \frac{(n-4)^2}{8} = \frac{8n-16}{8} = n - 2$. Since the difference between the two is an integer, they are both rounded the same way. Then all that remains is to verify that $B(i, m) - B(0, m - 1) = n - 2 = 4\,(m + 1) + i$, which is a matter of substitution. □

Remark 43. Note that A_1, which is the empty shape and has area 0, also corresponds to $\frac{(1+1)^2}{8} = \frac{1}{2}$ which we round to 1, therefore the claim does not hold for $n = 1$. However A_2, having area 1, gives $\frac{(2+1)^2}{8} = 1 + \frac{1}{8}$ and the closest integer to it is 1, as needed. This case, however, does not conveniently fall under the proof above.

5. Pareto optimality on the hexagonal grid

The characterization of isoperimetrically optimal shapes in the hexagonal tiling of $\mathbb{R}^2$ proceeds along the same lines as in the previously considered spaces. The first stage of analysis of the hexagon tiling is the same as seen above: the optimal shapes are shown to be simple, that is, intersections of six half planes each corresponding to one of the directions in the tiling. For reasons of space and similarity to previous sections, we omit this entirely, and only outline the pruning of simple shapes that are not optimal. For full details, please see our technical report [13].

The set of coordinates satisfying the six linear inequalities describing a simple shape, look in $\mathbb{Z}^2$ like a rectangle with (possibly zero) diagonals removed from two opposite corners (see Fig. 19). Shapes given in this form can be described using 4 integer parameters, and it is easy to calculate their area and perimeter size, allowing geometric constraints on optimal shapes to be expressed and proved in terms of those parameters. The number of cases to be considered is significantly reduced by assuming without loss of generality inequalities between the parameters, that geometrically translate into assumptions about the orientation of the shape in the hexagon tiling.

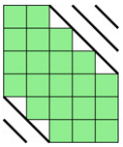

Fig. 19. The set of coordinates of a simple shape in the hexagon tiling can be described by the 4 numbers (6, 5, 2, 3) — its height, width, and the numbers of diagonals removed from the bottom left and top right corners respectively.

6. Concluding remarks

This paper completely characterized, for three common digital spaces the planar polyforms (shapes) that are optimal in having both the smallest perimeter for their area and the largest area for their perimeter. As a byproduct, the results generated interesting integer sequences, one of which was previously known only for a surprisingly different reason. We note that many variations in the definition of perimeter exist and may lead to different results. Some have been studied using a graph theory formulation, and some of this work is surveyed by Bezrukov and Serra [10]. The lack of regularity in some of the results illustrates the difficulties in dealing with discretizations of natural geometric concepts like the isoperimetric inequality. We discovered these first for X_4 [2], and have recently learned of a different approach to such analysis described in [9], which applies such results to achievement games. This is in addition to the applications to mathematical chemistry we have already mentioned, and those to swarm robotics [1] that originally motivated this work for us. Some advantages of the hexagonal grid and extensions to 3 dimensions are given by [3].

The Pareto optimality approach proposed here promotes the exploration of complete characterizations, in contrast with the construction of particular ones by spirals. This highlights surprises like the unconnected optimal shape of X_4 and its expansions which are optimal shapes not produced by spiral like constructions, and in other spaces allows us to state that the spiral produces all optimal shapes. Whether this approach is useful also for detailed explorations of higher dimensional discrete spaces is a topic for future work.

Acknowledgements

We thank the reviewers for reading our work in depth, and for comments that improved the paper. This research was partly supported by the Israel Ministry of Science Infrastructural grant No. 3-942.

References

[1] Yaniv Altshuler, Alfred M. Bruckstein, Israel A. Wagner, Swarm robotics for a dynamic cleaning problem, in: IEEE Swarm Intelligence Symposium 2005, SIS05, 2005, pp. 209–216.
[2] Yaniv Altshuler, Vladimir Yanovsky, Daniel Vainsencher, Israel A. Wagner, Alfred M. Bruckstein, On minimal perimeter polyminoes, Lecture Notes in Computer Science 4245 (2006) 17.
[3] Valentin E. Brimkov, Reneta P. Barneva, Analytical honeycomb geometry for raster and volume graphics, The Computer Journal 48 (2) (2005) 180–199.
[4] Alfred M. Bruckstein, The self-similarity of digital straight lines, in: International Conference on Pattern Recognition, 1990, pp. I: 485–490.
[5] Jon Brunvoll, Bjørg N. Cyvin, Sven J. Cyvin, More about extremal animals, Journal of Mathematical Chemistry (1993).
[6] Frank Harary, Heiko Harborth, Extremal animals, Journal of Combinatorics, Information and System Sciences 1 (1) (1976) 1–8.
[7] Benoit B. Mandelbrot, The Fractal Geometry of Nature, W.H. Freeman and Company, New York, San Francisco, 1983.
[8] Martin J. Osborne, Ariel Rubinstein, A Course in Game Theory, MIT Press, Cambridge, Mass, 1994.
[9] Nándor Sieben, Polyominoes with minimum site-perimeter and full set achievement games, European Journal of Combinatorics 29 (1) (2008) 108–117.
[10] S.L. Bezrukov, A local-global principle for vertex-isoperimetric problems, Discrete Mathematics 257 (25) (2002) 285–309.
[11] N. J. A. Sloane, The on-line encyclopedia of integer sequences, Published electronically at: http://www.research.att.com/~njas/sequences/. 1996–2006.
[12] Lawrence N. Stout, Two discrete forms of the Jordan curve theorem, The American Mathematical Monthly 95 (4) (1988) 332–336.
[13] Daniel Vainsencher, Alfred M. Bruckstein, Isoperimetrically optimal polyforms, CIS Technical Report 06, CS Department, Technion, 2007.
[14] Da-Lun Wang, Ping Wang, Discrete isoperimetric problems, SIAM Journal on Applied Mathematics 32 (4) (1977) 860–870.