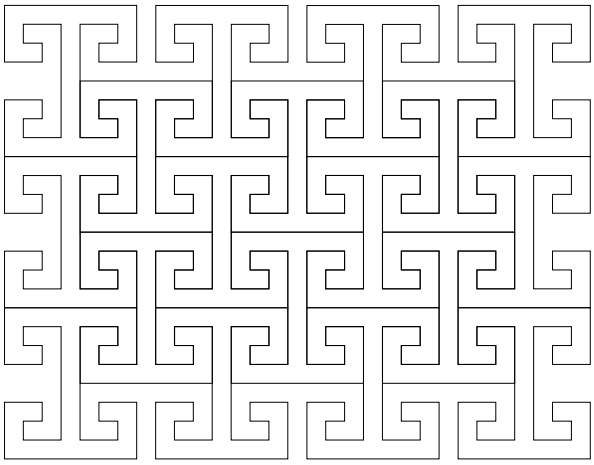
Tilings and Patterns



Fall 2025

by Jarkko Kari, University of Turku

Contents

1 Introduction **1**

2 Symmetries **1**
2.1 Isometries of the plane . . . 1
2.2 Fixed points . . . 5
2.3 Symmetries of a set of points . . . 7
2.4 Products of two reflections . . . 8
2.5 Parity . . . 10
2.6 Odd isometries . . . 12
2.7 Rosette groups . . . 13
2.8 Conjugacy . . . 16
2.9 Frieze groups . . . 18
2.10 Wallpaper groups . . . 21
2.11 Final remarks on discrete symmetry groups . . . 36

3 Tilings **37**
3.1 Basic definitions . . . 38
3.2 Tilings by regular polygons . . . 42

4 Wang tiles **48**
4.1 Periodic tilings . . . 49
4.2 Compactness principle . . . 51
4.3 Robinson's aperiodic tile set . . . 52
4.4 An aperiodic set of 14 Wang tiles . . . 56

5 Undecidable problems concerning tiles **61**
5.1 Turing machines . . . 63
5.2 The tiling problem with a seed tile . . . 65
5.3 Finite systems of forbidden patterns . . . 67
5.4 The periodic tiling problem . . . 69
5.5 The tiling problem . . . 73
5.6 The completion problem . . . 78
5.7 Beyond aperiodicity: arecursive tile sets . . . 81

6 Compact topology on Wang tilings **84**
6.1 Review of topology and metric spaces . . . 85
6.2 Basic facts about the configuration space . . . 91
6.3 Subshifts . . . 92
6.4 Orbits, transitivity and minimality . . . 93
6.5 Periodicity and recurrence properties . . . 95
6.6 Equicontinuity and isolated points . . . 98

7 A brief revisit to tilings by polygons **99**
7.1 Substitutions . . . 101
7.2 Amman's aperiodic tile set . . . 103
7.3 The extension and the periodicity theorems . . . 107
7.4 Hat: an aperiodic monotile . . . 108
7.5 Open problems . . . 119

1 Introduction

Informally, a tiling is a covering of the plane with tiles of various shapes in such a way that the tiles do not overlap each other. Often the tiles have simple shapes (e.g. polygons), and typically only a small number of different shapes are used in each tiling. Such tilings are everywhere around us: in pavements, quilt patterns, fabrics, brick walls, carpets, etc. Interest to decorative tilings is very old: Moors are an example of a culture that produced complex geometric patterns in tilings – famous examples can be found in the Alhambra at Granada, Spain.

In this course we learn about mathematical concepts relevant to tilings and patterns. The mathematical tools we use include high-school level geometry, elementary group theory, some topology, combinatorics and computation theory. After initial geometric considerations we work in detail on some computational questions on tilings, including decidability aspects. The basics of computation theory and other required material are provided during the course as needed, so that the course is made as self-contained as possible. In some instances we may rely on theorems from other fields that are presented without proofs, and in these instances an interested reader is directed to literature or other courses offered on these topics for more details and precise proofs.

2 Symmetries

Let us begin by investigating the fundamental concepts of symmetry.

2.1 Isometries of the plane

A plane isometry is any function $\alpha : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ that preserves distance:

$$\forall (x_1, y_1), (x_2, y_2) \in \mathbb{R}^2 : d(\alpha(x_1, y_1), \alpha(x_2, y_2)) = d((x_1, y_1), (x_2, y_2))$$

where the distance $d : \mathbb{R}^2 \times \mathbb{R}^2 \longrightarrow \mathbb{R}$ is the usual Euclidean distance defined by

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

In other words, α moves the points of the plane in a "rigid" motion that does not change any distances.

In these notes we'll denote points of the plane by capital letters, so the isometry property will be written as

$$\forall P, Q \in \mathbb{R}^2 : d(\alpha(P), \alpha(Q)) = d(P, Q).$$

Our first theorem states that an isometry is necessarily a bijection (that is, both one-to-one and onto). This implies that it has an inverse function. This inverse function is also an isometry.

Theorem 2.1 *An isometry is a bijection. Its inverse function is an isometry.*

Proof. Let α be an isometry. It is trivial that α is one-to-one (also the term "injective" is used). Namely, if $\alpha(P) = \alpha(Q)$ then

$$d(P, Q) = d(\alpha(P), \alpha(Q)) = 0,$$

which means that $P = Q$.

The proof that α is onto (also the term "surjective" is used) is more difficult, and is therefore left as a homework problem ;-)

Let $P, Q \in \mathbb{R}^2$ be arbitrary and denote $P' = \alpha^{-1}(P)$ and $Q' = \alpha^{-1}(Q)$. Then $P = \alpha(P')$ and $Q = \alpha(Q')$ so $d(P', Q') = d(P, Q)$, which proves that the inverse function α^{-1} preserves distance.

□

Our next observation states that isometries preserve shapes. More precisely, let us show that an isometry maps every line into a line, every triangle into a triangle (we say that it preserves lines and triangles), and the angle between two lines remains the same. Also betweenness and midpoints are preserved.

Theorem 2.2 *An isometry preserves lines, triangles, betweenness, midpoints, sizes of angles, and perpendicularity and parallelism of lines.*

Proof. Let α be an isometry. Let us prove the preservation of

- betweenness and midpoints: If three points P, Q and R are collinear, with point R between points P and Q, then $d(P,R) + d(R,Q) = d(P,Q)$. But then we have also

  $$d(P',R') + d(R',Q') = d(P',Q')$$

  where $P' = \alpha(P)$, $Q' = \alpha(Q)$ and $R' = \alpha(R)$. This means that points P', Q' and R' are also collinear, with R' between points P' and Q'. So betweenness is preserved.

  Since the inverse α^{-1} is also an isometry, the preservation works also in the inverse direction. In other words, R is between P and Q if and only if $\alpha(R)$ is between $\alpha(P)$ and $\alpha(Q)$.

  In the special case that R is the midpoint between P and Q we have that $d(P,R) = d(R,Q)$, so also $d(P',R') = d(R',Q')$, which means that R' is the midpoint between P' and Q'.

- triangles: Let $\triangle ABC$ be a triangle and, as usual, let us denote $A' = \alpha(A)$, $B' = \alpha(B)$ and $C' = \alpha(C)$. The triangle consists of those points P that are between A and B, between A and C, or between B and C. This is equivalent to $P' = \alpha(P)$ being between A' and B', between A' and C', or between B' and C'. Hence the image of triangle $\triangle ABC$ is the triangle $\triangle A'B'C'$.

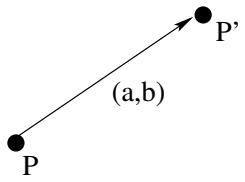
- lines: Let m be a line, and let A and B be two points on the line. Then the line consists exactly of those points P such that (i) P is between A and B, (ii) A is between B and P, or (iii) B is between A and P. This is equivalent to $P' = \alpha(P)$ being such that (i) P' is between A' and B', (ii) A' is between B' and P', or (iii) B' is between A' and P', where $A' = \alpha(A)$ and $B' = \alpha(B)$, which is equivalent to P' being on the line through points A' and B'.

- parallelism and perpendicularity of lines, as well as angles between lines: Take two different lines l and m. If they are parallel then they have no common points. Because α is one-to-one, their images $\alpha(l)$ and $\alpha(m)$ do not have any common points either, so they are parallel lines. Assume then that l and m are not parallel, in which case they intersect in one point P at some angle Θ. Let A and B be points of the lines l and m such that angle APB is of size Θ. Then the triangle $\triangle APB$ is congruent with its image $\triangle A'P'B'$ as the two triangles have same sides (SSS). Therefore the angle $A'P'B'$ is the same as the angle APB. In particular, l and m are perpendicular if and only if the angle is $90°$, so also perpendicularity is preserved.

□

The trivial isometry is the identity function ι that does not move any points: $\iota(P) = P$ for all $P \in \mathbb{R}^2$. Let us look into some non-trivial examples of isometries.

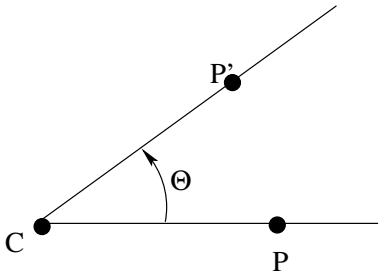
Example 1. Let $A = (a,b) \in \mathbb{R}^2$. A translation by vector $A = (a,b)$ shifts every point (x,y) into position $(x+a, y+b)$. We denote a translation by vector A as τ_A.



Every translation is clearly an isometry. Trivial translation $\tau_{(0,0)}$ is the trivial isometry ι.

□

Example 2. Let $C \in \mathbb{R}^2$ be a point, and $\Theta \in \mathbb{R}$ an angle. The rotation $\rho_{C,\Theta}$ by the (directed) angle Θ about C is the isometry that fixes point C, and otherwise takes point $P \neq C$ into the point P' where $d(C,P) = d(C,P')$ and Θ is the directed angle from CP to CP':

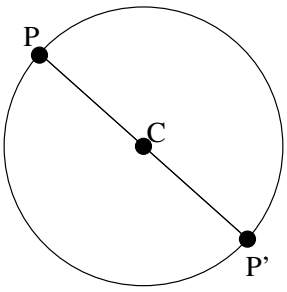


In terms of analytic geometry we say that point (x,y) is mapped to point (x',y') where

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\Theta & -\sin\Theta \\ \sin\Theta & \cos\Theta \end{pmatrix} \begin{pmatrix} x - c_x \\ y - c_y \end{pmatrix} + \begin{pmatrix} c_x \\ c_y \end{pmatrix}$$

where $C = (c_x, c_y)$. Point C is called the center of the rotation. The trivial rotation $\rho_{C,0}$ by the angle $0°$ is the trivial isometry ι.

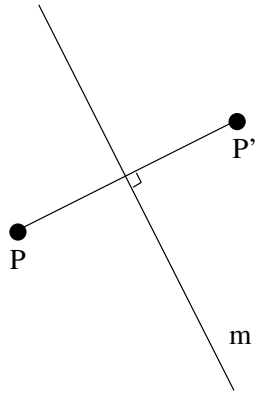
If $\Theta = 180°$ we get a special case of the rotation called the halfturn about point C, or the reflection in point C. Every point P is mapped to the point P' such that the center C is the midpoint between P and P':



Because halfturn about point C is an important particular case, we sometimes denote it by the special symbol σ_C.

□

Example 3. Let m be a line. The reflection σ_m in line m is the mapping that does not move the points of line m, but any point P outside line m is moved to the point P' such that line m is the perpendicular bisector of segment PP'.



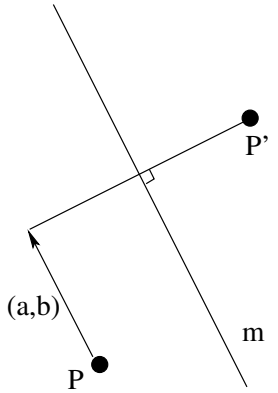
It follows immediately from the definition that $\sigma_m^{-1} = \sigma_m$, that is, the reflection σ_m is its own inverse. Isometries that are their own inverses are called involutions.

□

Example 4. A Glide reflection is a composition of a translation and a reflection in line m that is parallel with the direction of the translation. Let $A = (a, b) \in \mathbb{R}^2$ a vector of translation, and let m be a line parallel to A, that is,

$$m = \{(c, d) + t(a, b) \mid t \in \mathbb{R}\}$$

where (c, d) is some point of the line. The glide reflection $\gamma_{m,(a,b)}$ they specify reflects the points in line m and then translates them by vector A. In this particular case it does not matter in which order the two operations are performed: we may as well translate first and reflect later.



Line m is called the axis of the glide reflection. Notice that glide reflections with trivial translation vectors $A = (0, 0)$ are exactly the reflections.

□

Later we'll see that our four examples exhaust all possibilities: translations, rotations, reflections and glide reflections are the only isometries of the plane. (In fact, since reflection is a special type of glide reflection we can say that all isometries are translations, rotations or glide reflections.)

The composition $\alpha \circ \beta$ of two functions α and β is the function that first applies β to a point, and then applies α to the result, that is,

$$(\alpha \circ \beta)(x) = \alpha(\beta(x)).$$

If α and β are isometries then also their composition $\alpha \circ \beta$ is an isometry. Indeed, for any two points P and Q we have

$$d((\alpha \circ \beta)(P), (\alpha \circ \beta)(Q)) = d(\alpha(\beta(P)), \alpha(\beta(Q))) = d(\beta(P), \beta(Q)) = d(P, Q).$$

Function composition $\circ$ is an associative operation, and since the identity function ι and the inverses of all isometries are also isometries, we have the following theorem:

Theorem 2.3 *The set of plane isometries forms a group $\mathcal{I}$ under the operation of composition.*

□

We frequently drop the group operation sign "$\circ$" and simply write $\alpha\beta$ for $\alpha \circ \beta$. We then say that $\alpha\beta$ is the product of operations α and β. We also do not need to use parentheses in products as, because of the associativity, $\alpha(\beta\gamma) = (\alpha\beta)\gamma$. We simply write this as $\alpha\beta\gamma$. However, remember that the group of isometries is not commutative (=abelian) as in most cases $\alpha\beta \neq \beta\alpha$.

An element $\alpha \in \mathcal{I}$ is called an <u>involution</u> if $\alpha^2 = \iota$. Examples of involutions include all reflections in lines, as well as all halfturns. In fact, no other involutions exist. Review the following terms of group theory:

- generator set (=set of group elements such that every element of the group is a product of generators and their inverses),
- cyclic group (=a group that is generated by one element)
- order of a group (=number of elements. If the group contains an infinite number of elements then the group is called infinite, otherwise it is finite.)
- subgroup (=a subset of the group that is closed under the group operation and the operation of taking the inverse element. A subgroup itself is a group under the same group operation)
- cancellation laws ($\alpha\beta = \alpha\gamma$ implies $\beta = \gamma$, and $\beta\alpha = \gamma\alpha$ implies $\beta = \gamma$.)

In the rest of this chapter we try to understand the structure of the group $\mathcal{I}$. We want to show that our examples exhaust all possibilities, and to find out how the group operation combines these isometries.

2.2 Fixed points

The two main results of this section are the following:

1. To verify that two given isometries α and β are the same, it is sufficient to verify that they agree on some three points that are not collinear (Corollary 2.6).
2. Every isometry is a product of at most three reflections (Corollary 2.7).

We say that P is a <u>fixed point</u> of isometry α if $\alpha(P) = P$. We also say that α <u>fixes</u> point P.

Lemma 2.4 *If an isometry α fixes two distinct points P and Q, then it fixes every point of the line m that contains P and Q.*

Proof. Assume that α fixes points P and Q of line m, and let R be any point of the line m. Because α preserves lines, $\alpha(R)$ is on the same line with $\alpha(P) = P$ and $\alpha(Q) = Q$, that is, $\alpha(R)$ is on line m. Because $d(\alpha(R), P) = d(R, P)$ and $d(\alpha(R), Q) = d(R, Q)$, we must have $\alpha(R) = R$. (There are two points at distance $d(R, P)$ from P, and these two points have different distances from point Q. So only one of these two points can have distance $d(R, Q)$ from Q, namely point R.)

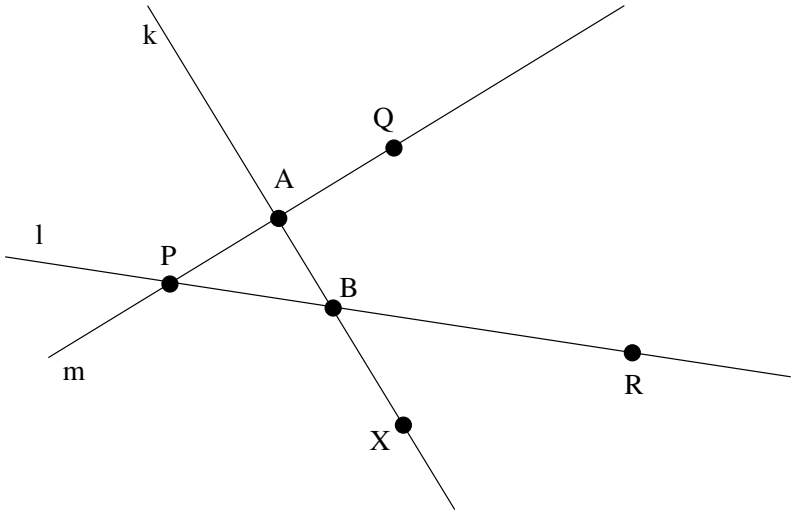
□

Consider three points P, Q and R that are not collinear, i.e. are not on the same line. As a corollary of the next theorem we get that their images $\alpha(P), \alpha(Q)$ and $\alpha(R)$ uniquely determines the isometry α. We also prove that every isometry is a product of at most three reflections.

Theorem 2.5 *Let α be an isometry.*

1. *If α fixes three non-collinear points, then $\alpha = \iota$.*
2. *If α fixes two points then $\alpha = \iota$ or α is a reflection.*
3. *If α fixes exactly one point then α is a product of two reflections.*

Proof. 1. Assume that α fixes three non-collinear points P, Q and R. Let m and l be the lines that contain P and Q, and P and R, respectively. According to Lemma 2.4, α fixes all points that belong to lines m or l. Let X be an arbitrary point outside lines m and l. There exists a line k that goes through X and intersects m and l at distinct points A and B.



Because α fixes A and B then, according to Lemma 2.4, it also fixes all points of line k, which means that it also fixes point X. As X was an arbitrary point, we conclude that α fixes all points of the plane, so $\alpha = \iota$.

2. Assume then that α fixes two distinct points P and Q, and suppose that $\alpha \neq \iota$. Then there exists some point R such that $\alpha(R) \neq R$. Notice that P, Q and R cannot be collinear (Lemma 2.4). Denote $R' = \alpha(R)$, and let m be the perpendicular bisector of the segment RR'. Then $R' = \sigma_m(R)$ where σ_m is the reflection in line m. Because $d(R', P) = d(R, P)$ and $d(R', Q) = d(R, Q)$, points P and Q are on the perpendicular bisector m. We have $\sigma_m(P) = P$ and $\sigma_m(Q) = Q$. The isometry $\sigma_m^{-1}\alpha$ hence fixes three non-collinear points P, Q and R so, according to case 1 of the theorem, $\sigma_m^{-1}\alpha = \iota$. This proves that $\alpha = \sigma_m$ is a reflection.

3. Assume that isometry α fixes exactly one point P. Let Q a different point, so $Q' = \alpha(Q)$ is different from Q. Let l be the perpendicular bisector of the segment QQ'. Triangle $\triangle QPQ'$ is isosceles, so the point P is on the line l. Then $\sigma_l^{-1}\alpha$ fixes two points P and Q, so according to case 2 either $\sigma_l^{-1}\alpha = \iota$ or $\sigma_l^{-1}\alpha = \sigma_m$ for some line m. The first alternative $\alpha = \sigma_l$ is not possible because then α fixes more than one point – it fixes all points of line l. So we must have the second alternative $\alpha = \sigma_l\sigma_m$.

□

Corollary 2.6 *If α and β are two isometries such that $\alpha(P) = \beta(P)$, $\alpha(Q) = \beta(Q)$ and $\alpha(R) = \beta(R)$, and points P, Q and R are not collinear, then $\alpha = \beta$.*

Proof. Isometry $\alpha^{-1}\beta$ fixes non-collinear points P, Q and R, so $\alpha^{-1}\beta = \iota$. This implies $\alpha = \beta$.

□

Corollary 2.7 *Every isometry is a product of at most three reflections.*

Proof. If α fixes at least one point then, according to the theorem, α is a product of at most two reflections. Assume then that α does not fix any points. Let P be an arbitrary point, and let m be the perpendicular bisector of the segment $P\alpha(P)$. Then $\sigma_m^{-1}\alpha$ fixes point P, so $\sigma_m^{-1}\alpha$ is a product of at most two reflections and, therefore, α is a product of at most three reflections.

□

The proofs provide a simple method of finding the reflections when we know the images $P_0 = \alpha(P), Q_0 = \alpha(Q)$ and $R_0 = \alpha(R)$ of three given non-collinear points P, Q and R. We simply find reflections that match the points one-by-one:

1. If $P \neq P_0$ then we first reflect in line m that is the perpendicular bisector of the segment PP_0. This maps P to its correct position P_0. Let Q' and R' be the images of Q and R under the first reflection.

2. If $Q' \neq Q_0$ then we reflect in line l that is the perpendicular bisector of the segment $Q'Q_0$. Notice that point P_0 is on this bisector because $d(P_0, Q_0) = d(P, Q) = d(P_0, Q')$. After the second reflection, points P and Q have been mapped to their correct positions P_0 and Q_0. Let R'' be the image of R after the first two reflections.

3. If $R'' \neq R_0$ then we finally reflect in line k that is the perpendicular bisector of R'' and R_0. It is easy to see that P_0 and Q_0 are on this bisector:

$$d(P_0, R_0) = d(P, R) = d(P_0, R'') \qquad \text{and} \qquad d(Q_0, R_0) = d(Q, R) = d(Q_0, R'').$$

After steps 1–3, points P, Q and R have been mapped in their correct positions P_0, Q_0 and R_0

2.3 Symmetries of a set of points

Let $s \subseteq \mathbb{R}^2$ be a set of points. We say that isometry α is a symmetry of set s iff $\alpha(s) = s$.

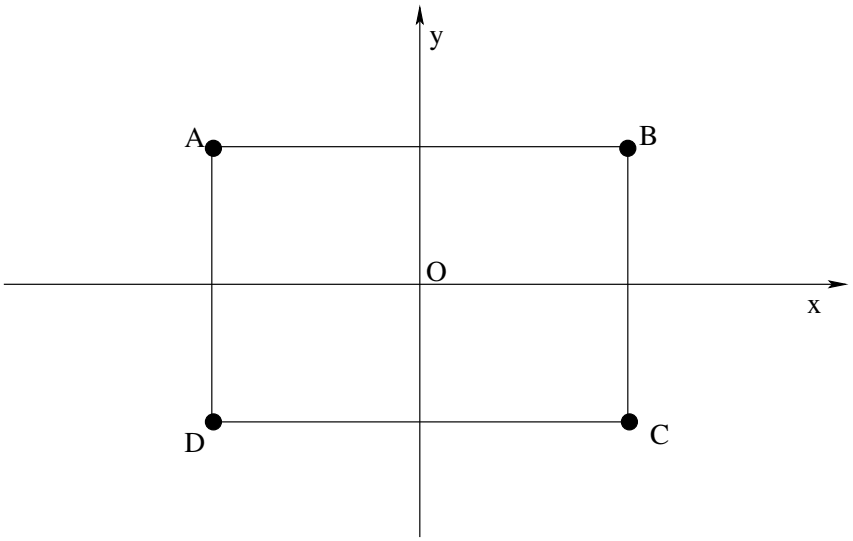
Theorem 2.8 *Let $s \subseteq \mathbb{R}^2$ be arbitrary. The symmetries of s form a subgroup of $\mathcal{I}$, the group of isometries.*

Proof. Every set has at least one symmetry, namely the trivial isometry ι. If $\alpha(s) = s$ then $\alpha^{-1}(s) = \alpha^{-1}(\alpha(s)) = s$, so the inverse of each symmetry of s is also a symmetry of s. Let α and β be two symmetries of s. Then $\alpha\beta(s) = \alpha(s) = s$ so the product $\alpha\beta$ is also a symmetry of s.

□

The set of symmetries of s is called the symmetry group of s. Notice that $\mathcal{I}$ itself is the symmetry group of $s = \mathbb{R}^2$.

Example 5. Let s be a rectangle $ABCD$ that is not a square. Let us position s in such a way that its center is at the origin $(0, 0)$, and its sides are parallel to the x- and y-axes.



Any symmetry of s must permute the corners of the rectangle. Corner A may be mapped into any of the four corners A, B, C and D, after which the images of the other corners B, C and D are uniquely determined. We proved in the previous section that three non-collinear points A, B and C determine the entire isometry (Corollary 2.6), so the symmetry group s contains exactly four symmetries. These are ι, two reflections σ_h and σ_v in the x- and y-axes, and the halfturn σ_O about the origin O. These form Klein's Vierergruppe V_4.

□

Example 6. If s is a square $ABCD$ then its symmetry group contains eight elements, so a square is "more" symmetric than a non-square rectangle. In the square we may map the corner A into any of the four corners, after which corner B has still two possible images. Then the images of C and D are uniquely determined.

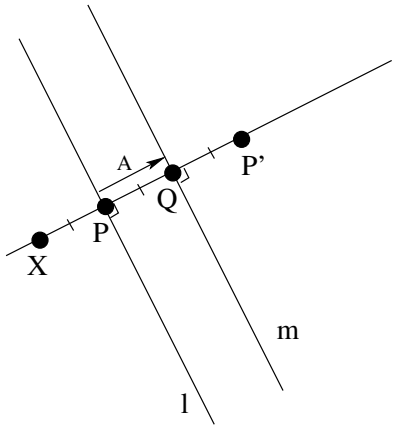
□

2.4 Products of two reflections

We know that every isometry is a product of at most three reflections. In order to characterize all isometries we need to investigate the products of two or three reflections. Let us start by products of two reflections.

Theorem 2.9 *The product of two reflections in parallel lines m and l is a translation in the direction perpendicular to l and m by a distance that is twice the distance from l to m. Conversely, every translation is a product of two reflections in parallel lines, both perpendicular to the direction of the translation. One of the lines can be chosen freely (as long as it is perpendicular to to the translation).*

Proof. Let m and l be two parallel lines. If $m = l$ then $\sigma_m\sigma_l = \iota = \tau_{(0,0)}$. Assume then that $m \neq l$. Let A be the vector from l to m that is perpendicular to m and l. To prove that $\sigma_m\sigma_l = \tau_{2A}$ it is enough to show that $\sigma_m\sigma_l(P) = \tau_{2A}(P)$ for every point P of line l, and that $\sigma_m\sigma_l(X) = \tau_{2A}(X)$ for some point X outside of line l. Then the result follows from Corollary 2.6.



Referring to the figure above, we have that for every $P \in l$

$$\sigma_m\sigma_l(P) = \sigma_m(P) = P' = \tau_{2A}(P).$$

Analogously, by reversing the roles of lines m and l, we have that for an arbitrary $Q \in m$

$$\sigma_l\sigma_m(Q) = \tau_{-2A}(Q).$$

Let $X = \sigma_l\sigma_m(Q) = \tau_{-2A}(Q)$. Then X is not on line l, and

$$\sigma_m\sigma_l(X) = \sigma_m\sigma_l\sigma_l\sigma_m(Q) = Q = \tau_{2A}\tau_{-2A}(Q) = \tau_{2A}(X).$$

The second part of the theorem follows directly from the first part: Let τ be a non-trivial translation, and let P be an arbitrary point and $P' = \tau(P)$. Let l and m be the lines perpendicular to the segment PP' through P and the midpoint of PP', respectively. Then, according to the first part, $\sigma_m\sigma_l = \tau$.

□

Corollary 2.10 *The product of three reflections in three parallel lines is a reflection in a parallel line.*

Proof. Let l, m and n be any three parallel lines. Let p be a fourth parallel line whose distance from line n is the same as the distance of line l from line m. Then $\sigma_l\sigma_m$ and $\sigma_p\sigma_n$ are the same translation. Multiplying by σ_n from the right gives $\sigma_l\sigma_m\sigma_n = \sigma_p$.

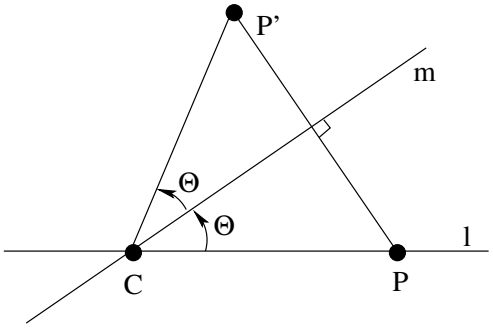
□

Consider then two reflections in lines that are not parallel:

Theorem 2.11 *The product of two reflections in intersecting lines is a rotation about the point of intersection, and the angle of the rotation is twice the angle between the lines. Conversely, every rotation about point C is a product of two reflections in lines through point C. One of these lines can be chosen freely.*

Proof. Let l and m be lines that intersect at point C. Let Θ be the directed angle between them measured from l to m. Let us prove that $\sigma_m\sigma_l = \rho_{C,2\Theta}$ by showing that $\sigma_m\sigma_l$ and $\rho_{C,2\Theta}$ agree on three non-collinear points: all points of line l, and one point X that is outside of line l.

First, as all the isometries σ_m, σ_l and $\rho_{C,2\Theta}$ fix point C, we have $\sigma_m\sigma_l(C) = C = \rho_{C,2\Theta}(C)$. Let $P \neq C$ be a point on line l, and let $P' = \rho_{C,2\Theta}(P)$. Line m is the perpendicular bisector of PP', so $P' = \sigma_m(P) = \sigma_m\sigma_l(P)$.



So far we have proved that $\sigma_m\sigma_l(P) = \rho_{C,2\Theta}(P)$ for all $P \in l$. Analogously, by reversing the roles of lines m and l, we have that $\sigma_l\sigma_m(Q) = \rho_{C,-2\Theta}(Q)$ for an arbitrary point $Q \neq C$ of line m. Denote $X = \sigma_l\sigma_m(Q) = \rho_{C,-2\Theta}(Q)$. Then X is not on line l and

$$\sigma_m\sigma_l(X) = \sigma_m\sigma_l\sigma_l\sigma_m(Q) = Q = \rho_{C,2\Theta}\rho_{C,-2\Theta}(Q) = \rho_{C,2\Theta}(X).$$

To prove the second part of the theorem, consider an arbitrary rotation $\rho_{C,\Theta}$. Let l be an arbitrary line through the center C of the rotation, and let m be the line through point C that meets line l in the directed angle $\Theta/2$. According to the first part of the theorem we have $\sigma_m\sigma_l = \rho_{C,\Theta}$.

□

Corollary 2.12 *Halfturn σ_C is the product of two reflections in any two perpendicular lines through C. In particular, reflections in perpendicular lines commute.*

□

Corollary 2.13 *The product of three reflections in lines through the common point C is a reflection in a line through point C.*

Proof. As in the proof of Corollary 2.10, let l, m and n be any three lines through point C. Let p be a fourth line through C that forms with line n the same angle as line l forms with line m. Then $\sigma_l\sigma_m$ and $\sigma_p\sigma_n$ are the same rotation about point C. Multiplying by σ_n from the right gives $\sigma_l\sigma_m\sigma_n = \sigma_p$.

□

2.5 Parity

As we proved previously, all isometries are products of some reflections, in fact, of at most three reflections. The representation of an isometry as a product of reflections is, however, not unique. For example, we can always add $\sigma_m\sigma_m$ to the end of any sequence of reflections, thus increasing the number of reflections in the sequence by two. However, it turns out that the parity of the number of reflections is always the same. We call isometry α <u>even</u> if it is a product of an even number of reflections, and <u>odd</u> if it is a product of an odd number of reflections. Next we want to show that no isometry can be both even and odd at the same time, that is, even and odd products of reflections can never be equal.

First we can make the following easy observation: A product of two reflections is not a reflection. Indeed, we know from the results of the previous section that a product of two reflections is either a translation or a rotation. Translations have no fixed points, rotations have exactly one fixed point, and the trivial isometry ι fixes all points. In contrast, the fixed points of a reflection form a line. So $\sigma_m\sigma_l \neq \sigma_k$ for all lines m, l and k.

The following theorem provides a method of reducing by two the number of terms in any long product of reflections:

Theorem 2.14 *A product of four reflections is a product of two reflections.*

Proof. We use the following lemma twice:

Lemma 2.15 *If m and l are two lines and P is a point, then there are lines p and q such that $\sigma_m\sigma_l = \sigma_p\sigma_q$, and line q contains point P.*

Proof of the lemma. If m and l are parallel, then we choose as q the line that is parallel to m and l and goes through point P. By corollary 2.10 we have $\sigma_m\sigma_l\sigma_q = \sigma_p$ for some line p, so $\sigma_m\sigma_l = \sigma_p\sigma_q$.

If m and l intersect at some point Q, then we choose as q a line through points P and Q. By corollary 2.13 we have $\sigma_m\sigma_l\sigma_q = \sigma_p$ for some line p, so $\sigma_m\sigma_l = \sigma_p\sigma_q$.

□

Consider a product $\sigma_m\sigma_l\sigma_k\sigma_n$ of four reflections. Let P be an arbitrary point on line n. According to the lemma, $\sigma_l\sigma_k = \sigma_p\sigma_q$ where line q contains point P. Then we apply the lemma again: $\sigma_m\sigma_p = \sigma_r\sigma_s$ where s contains point P. We have

$$\sigma_m\sigma_l\sigma_k\sigma_n = \sigma_m\sigma_p\sigma_q\sigma_n = \sigma_r\sigma_s\sigma_q\sigma_n,$$

and lines n, q and s go through point P. Then, by Corollary 2.13 the product $\sigma_s\sigma_q\sigma_n = \sigma_t$ for some line t. Hence

$$\sigma_m\sigma_l\sigma_k\sigma_n = \sigma_r\sigma_t.$$

□

Corollary 2.16 *A product of three reflections cannot equal a product of two reflections.*

Proof. Assume that

$$\sigma_m\sigma_l\sigma_k = \sigma_n\sigma_r.$$

Multiplying from left by σ_n gives

$$\sigma_n\sigma_m\sigma_l\sigma_k = \sigma_r.$$

According to the theorem there exist lines p and q such that

$$\sigma_n\sigma_m\sigma_l\sigma_k = \sigma_p\sigma_q,$$

so $\sigma_p\sigma_q = \sigma_r$, a contradiction.

□

Corollary 2.17 *A product of an even number of reflections cannot equal a product of an odd number of reflections.*

Proof. By using the theorem we can reduce by two the number of reflections in any product of at least four reflections. In this way, any even length sequence can be reduced into a product of two reflections, and any odd length sequence reduces into a length one or a length three sequence. As a product of two reflections cannot equal a product of one or three reflections, we have the desired result.

□

Now we know that every isometry is either even or odd, but not both. Notice that odd isometries correspond to "flipping" the plane over, turning all shapes into their mirror images. As every even isometry is a product of two reflections, we have

Theorem 2.18 *Even isometries are exactly the translations and the rotations.*

□

Notice also that even isometries form a subgroup of $\mathcal{I}$. Indeed, the inverse of the even isometry $\sigma_m\sigma_l$ is the even isometry $\sigma_l\sigma_m$, and the product of two even isometries $\sigma_m\sigma_l$ and $\sigma_n\sigma_p$ is the even isometry $\sigma_m\sigma_l\sigma_n\sigma_p$. Let us denote the group of even isometries by $\mathcal{E}$.

2.6 Odd isometries

Let's turn our attention to odd isometries. The goal of this section is to prove that every odd isometry is a glide reflection (where we understand that a plain reflection is also a glide reflection with a zero glide.) Recall that we use the notation σ_P for the halfturn about point P.

Lemma 2.19 *Isometry α is a glide reflection if and only if $\alpha = \sigma_P\sigma_l$ for some point P and line l. This is also equivalent to $\alpha = \sigma_k\sigma_Q$ for some line k and point Q.*

Proof. Let α be a glide reflection. By definition, $\alpha = \sigma_m\tau_A$ where the translation τ_A is in the direction of line m. By Theorem 2.9 $\tau_A = \sigma_k\sigma_l$ where lines k and l are perpendicular to line m. We have $\alpha = \sigma_m\sigma_k\sigma_l$. Corollary 2.12 states that the product $\sigma_m\sigma_k$ of two reflections in perpendicular lines is the halfturn σ_P about the intersection point P of lines m and k. We have

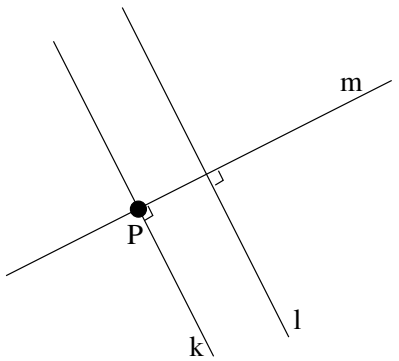
$$\alpha = \sigma_m\sigma_k\sigma_l = \sigma_P\sigma_l$$

as desired. We also have $\sigma_P = \sigma_k\sigma_m$, so

$$\alpha = \sigma_P\sigma_l = \sigma_k\sigma_m\sigma_l = \sigma_k\sigma_Q$$

where Q is the point where perpendicular lines m and l intersect.

For the converse claim, assume that α is the isometry $\sigma_P\sigma_l$ for some point P and line l. Let k be the line through point P that is parallel to line l, and let m be the line through point P that is perpendicular to lines k and l. Then, by Corollary 2.12, $\sigma_P = \sigma_m\sigma_k$.



We have

$$\alpha = \sigma_P\sigma_l = \sigma_m\sigma_k\sigma_l = \sigma_m\tau_A$$

where τ_A is in the direction of line m. Hence α is a glide reflection.

Analogously, if $\alpha = \sigma_k\sigma_Q$, and lines m and l go through point Q, and l is parallel and m perpendicular to k, then

$$\alpha = \sigma_k\sigma_Q = \sigma_k\sigma_m\sigma_l = \sigma_m\sigma_k\sigma_l = \sigma_m\tau_A$$

where A is in the direction of line m.

□

Now we are able to prove the main result on odd isometries:

Theorem 2.20 *Every odd isometry is a glide reflection.*

Proof. Let α be an odd isometry. Then it is either a reflection (which is a special type of a glide reflection) or a product of three reflections. Let $\alpha = \sigma_m\sigma_l\sigma_k$. Let P be an arbitrary point on line k. By Lemma 2.15 there exist lines p and q such that $\sigma_m\sigma_l = \sigma_p\sigma_q$ and line q goes through point P. We have

$$\alpha = \sigma_m\sigma_l\sigma_k = \sigma_p\sigma_q\sigma_k,$$

and $P \in k, q$. Let n be the line through point P that is perpendicular to line p. As lines n, q and k all go through point P, the product $\sigma_n\sigma_q\sigma_k$ is some reflection σ_r, see Corollary 2.13. Then $\sigma_q\sigma_k = \sigma_n\sigma_r$, and

$$\alpha = \sigma_p\sigma_n\sigma_r.$$

Lines n and p are perpendicular, so the product $\sigma_p\sigma_n$ is a halfturn σ_Q, where Q is the point where n and p intersect. We have

$$\alpha = \sigma_Q\sigma_r,$$

and it now follows from Lemma 2.19 that α is a glide reflection.

□

Now we have classified all isometries of the plane. Even isometries are translations and rotations, and odd isometries are glide reflections (including reflections without glides).

2.7 Rosette groups

Rosette groups are the finite subgroups of $\mathcal{I}$. In this section we prove that the rosette groups are the cyclic groups C_n and the dihedral groups D_n, for $n \geq 1$, defined as follows:

The cyclic group C_n consists of n rotations about the same center P. It is generated by the single rotation $\rho = \rho_{P,\frac{360^\circ}{n}}$, so the elements of C_n are $\rho, \rho^2, \ldots, \rho^n = \iota$. Notice that strictly speaking there are infinitely many groups C_n as the center P can be any point of the plane, but they are all obviously isomorphic with each other.

The dihedral group D_n includes C_n, and in addition it contains reflections in n lines that meet at P (the center of the rotations) at angles that are multiples of $\frac{360^\circ}{2n}$. Notice that the composition of two such reflections is a rotation that belongs to C_n. There are $2n$ elements in D_n: namely n rotations $\rho, \rho^2, \ldots, \rho^n = \iota$, and n reflections that can be expressed as $\rho\sigma, \rho^2\sigma, \ldots, \rho^n\sigma = \sigma$, where σ is any one of the reflections.

Here are the cases with small $n = 1$ and 2:

- $C_1 = \{\iota\}$ and $D_1 = \{\iota, \sigma_m\}$,
- $C_2 = \{\iota, \sigma_P\}$ and $D_2 = \{\iota, \sigma_P, \sigma_m, \sigma_l\}$, where m and l are perpendicular lines through point P.

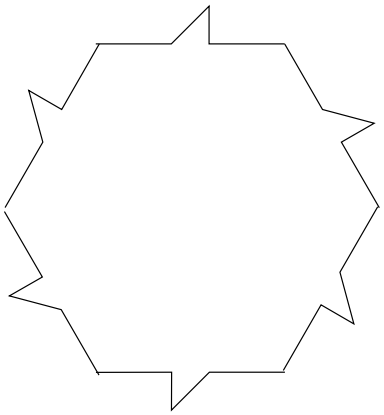
Example 7. The symmetry group of a polygon with n edges and vertices (called n-gon) can contain at most $2n$ elements. Indeed, any symmetry α must map vertices into vertices, and neighboring vertices into neighboring vertices. Fixed vertex A has at most n possible images. Adjacent vertex B then has at most two alternatives as it must be mapped into one of the two vertices next to $\alpha(A)$. After this, the symmetry is uniquely determined.

Let us show that the regular n-gon has exactly $2n$ symmetries, and the symmetry group is the dihedral group D_n. Let P be the center of the regular n-gon. It is clear that the rotation $\rho = \rho_{P,\frac{360^\circ}{n}}$ is a symmetry of the n-gon. If m is line through P and one of the vertices then also $\sigma = \sigma_m$ is a symmetry. As the symmetries form a group, all products of σ and ρ are symmetries. These include the n rotations

$\rho, \rho^2, \ldots, \rho^n = \iota$ generated by ρ, and n distinct odd isometries $\rho\sigma, \rho^2\sigma, \ldots, \rho^n\sigma = \sigma$. (These are distinct as $\rho^i\sigma = \rho^j\sigma \Longrightarrow \rho^i = \rho^j$.) These are exactly the elements of D_n. There can be no other isometries as no n-gon can have more than $2n$ symmetries.

□

Example 8. Cyclic group C_n is the symmetry group of a polygon that is obtained from a regular n-gon by replacing each edge with a "directed edge", for example as follows:



□

Before proving that no other finite subgroups of $\mathcal{I}$ exists, let us first figure out multiplication rules of even isometries.

Theorem 2.21 *1. The product of two translations is a translation.*

2. *A rotation by angle Θ followed by a rotation by angle Φ is a rotation by angle $\Theta + \Phi$, unless $\Theta + \Phi$ is a multiple of $360°$, in which case the product is a translation.*

3. *A translation followed by a non-trivial rotation by Θ is a rotation by Θ. Also, a non-trivial rotation by Θ followed by a translation is a rotation by Θ.*

Proof.

1. Trivial: it follows from the definition of translations that $\tau_A\tau_B = \tau_{A+B}$.

2. If the two rotations are about the same center P then the claim is trivial: $\rho_{P,\Theta}\rho_{P,\Phi} = \rho_{P,\Theta+\Phi}$. Assume then that the two rotations are about different points A and B. Let m be the line through points A and B. According to Theorem 2.11 there exist lines l and n through points A and B, respectively, such that

$$\rho_{A,\Theta} = \sigma_m\sigma_l \qquad \text{and} \qquad \rho_{B,\Phi} = \sigma_n\sigma_m,$$

so

$$\rho_{B,\Phi}\rho_{A,\Theta} = \sigma_n\sigma_m\sigma_m\sigma_l = \sigma_n\sigma_l.$$

Moreover, the directed angle from l to m is $\Theta/2$ and the directed angle from m to n is $\Phi/2$, so the directed angle from l to n is $(\Theta + \Phi)/2$. If this angle is a multiple of $180°$ then lines l and n are parallel, that is, if $\Theta + \Phi$ is a multiple of $360°$ then $\rho_{B,\Phi}\rho_{A,\Theta}$ is a translation. Otherwise lines l and n are not parallel, so $\rho_{B,\Phi}\rho_{A,\Theta}$ is a rotation by angle $\Theta + \Phi$.

3. Let τ be a translation and ρ a non-trivial rotation by angle Θ. Then $\tau = \sigma_l\sigma_m$ for parallel lines l and m, and $\rho = \sigma_n\sigma_k$ where the angle from line k to line n is $\Theta/2$. By Theorem 2.11 We can choose k to be parallel to l and m. Then $\rho\tau = \sigma_n\sigma_k\sigma_l\sigma_m$. Because k, l and m are parallel lines, by Corollary 2.10 the product $\sigma_k\sigma_l\sigma_m$ is a reflection σ_p where p is also parallel to k, l and m. The angle from line p to line n is $\Theta/2$, so $\rho\tau = \sigma_n\sigma_p$ is a rotation by angle Θ.

   Analogously, we could have chosen k and n so that n is parallel to l and m, in which case $\sigma_l\sigma_m\sigma_n = \sigma_q$ for a line q in the same direction. Then

$$\tau\rho = \sigma_l\sigma_m\sigma_n\sigma_k = \sigma_q\sigma_k$$

   is a rotation by angle Θ.

□

By iterating the theorem we easily get a rule for composing an arbitrary number of rotations:

$$\rho_{C_1,\Theta_1} \circ \rho_{C_2,\Theta_2} \circ \cdots \circ \rho_{C_n,\Theta_n}$$

is a rotation by angle $\Theta = \Theta_1 + \Theta_2 + \cdots + \Theta_n$, unless Θ is a multiple of 360°, in which case the product is a translation.

Corollary 2.22 *If a subgroup of $\mathcal{I}$ contains two non-trivial rotations about different centers then it also contains a non-trivial translation*

Proof. Let $\rho_{A,\Theta}$ and $\rho_{B,\Phi}$ be two non-trivial rotations and $A \neq B$. According to our theorem

$$\rho_{B,\Phi}^{-1}\rho_{A,\Theta}^{-1}\rho_{B,\Phi}\rho_{A,\Theta} = \rho_{B,-\Phi}\rho_{A,-\Theta}\rho_{B,\Phi}\rho_{A,\Theta}$$

is a translation. If it were the trivial translation ι then

$$\rho_{A,\Theta}\rho_{B,\Phi} = \rho_{B,\Phi}\rho_{A,\Theta}$$

but this is not possible as it was proved in a homework problem that non-trivial rotations about different centers do not commute.

□

Now we are ready to prove the result mentioned in the beginning of this section:

Theorem 2.23 (Leonardo da Vinci's Theorem) *A finite subgroup of $\mathcal{I}$ is either a cyclic group C_n or a dihedral group D_n.*

Proof. Let H be a finite subgroup of $\mathcal{I}$. Every non-trivial translation generates an infinite subgroup, so H cannot contain non-trivial translations. If γ is a glide reflection with glide vector A then γ^2 is a translation by vector $2A$, so H cannot contain any glide reflections except plain reflections. So only rotations and reflections are possible.

By the previous lemma, all rotations in H must have the same center P. Let $\rho = \rho_{P,\Theta}$ be the rotation having the smallest positive angle Θ among all rotations in H. It exists as H is finite. Let $\rho_{P,\Phi} \in H$. For every real number Φ there exists an integer k such that $0 \leq \Phi - k\Theta < \Theta$. Because the rotation by $\Phi - k\Theta$ is in H, and because Θ is the smallest positive angle, we must have $\Phi - k\Theta = 0$. This means that $\rho_{P,\Phi} = \rho^k$. We have proved that ρ generates the rotations of H. This means that the set of even isometries in H is $\{\rho, \rho^2, \ldots, \rho^n\} = C_n$ for some n.

If there are no reflections in H then $H = C_n$. Assume then that there is at least one reflection σ in H. Then there are at least n distinct odd isometries $\sigma\rho, \sigma\rho^2, \ldots, \sigma\rho^n$ in H. On the other hand, if $\alpha \in H$ is odd then $\sigma\alpha$ is even, so $\sigma\alpha = \rho^k$ for some $k = 1, 2, \ldots, n$. This means that $\alpha = \sigma\rho^k$, and we have proved that $H = D_n$.

□

Corollary 2.24 *The symmetry group of every polygon is a cyclic group or a dihedral group.*

Proof. In the example at the beginning of the section we concluded that the symmetry group of an n-gon contains at most $2n$ elements, so it is finite.

□

2.8 Conjugacy

Two elements x and y of a group G are called conjugate if there exists an element $\alpha \in G$ such that $x = \alpha y \alpha^{-1}$. It is easy to see that conjugacy is an equivalence relation. Its equivalence classes are called the conjugacy classes of the group.

It turns out that in the group $\mathcal{I}$ conjugate isometries are of the same type (both translations, both rotations, both reflections or both glide reflections):

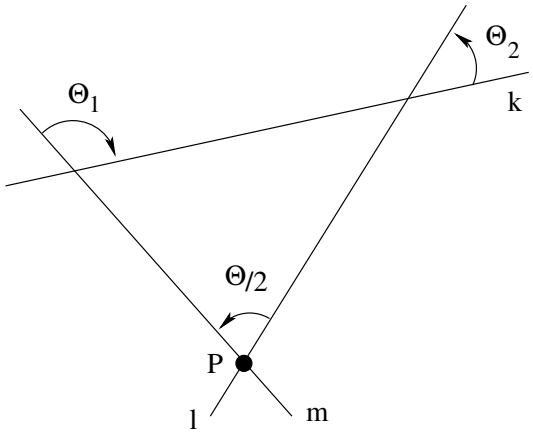
Theorem 2.25 *Let $\alpha \in \mathcal{I}$ be an arbitrary isometry.*

1. *Let $\sigma = \sigma_m$ be the reflection in line m. Then $\alpha\sigma\alpha^{-1}$ is the reflection $\sigma_{\alpha(m)}$ in line $\alpha(m)$.*
2. *Let $\tau = \tau_{B-A}$ be the translation that moves point A to point $B = \tau(A)$. Then $\alpha\tau\alpha^{-1}$ is the translation $\tau_{\alpha(B)-\alpha(A)}$ that moves point $\alpha(A)$ to point $\alpha(B)$.*
3. *Let $\rho = \rho_{P,\Theta}$ be a rotation about point P. Then $\alpha\rho\alpha^{-1}$ is the rotation $\rho_{\alpha(P),\pm\Theta}$ about point $\alpha(P)$, where the angle is $+\Theta$ if α is even, and $-\Theta$ if α is odd.*
4. *Let $\gamma = \gamma_{m,B-A}$ be a glide reflection. Then $\alpha\gamma\alpha^{-1}$ is the glide reflection $\gamma_{\alpha(m),\alpha(B)-\alpha(A)}$.*

Proof.

1. Isometry $\alpha\sigma_m\alpha^{-1}$ is an odd isometry that fixes every point $\alpha(P)$ of line $\alpha(m)$. The only odd isometry with this property is the reflection in line $\alpha(m)$.
2. Let $\tau = \tau_{B-A}$ be the translation that moves A to B. Then $\tau = \sigma_m\sigma_l$ for two parallel lines m and l. According to case 1 above, $\alpha\sigma_m\alpha^{-1} = \sigma_{\alpha(m)}$ and $\alpha\sigma_l\alpha^{-1} = \sigma_{\alpha(l)}$. We get
$$\alpha\tau\alpha^{-1} = \alpha\sigma_m\sigma_l\alpha^{-1} = \alpha\sigma_m\alpha^{-1}\alpha\sigma_l\alpha^{-1} = \sigma_{\alpha(m)}\sigma_{\alpha(l)}.$$
Isometries preserve parallelism of lines, so $\alpha(m)$ and $\alpha(l)$ are parallel lines, which means that $\alpha\tau\alpha^{-1}$ is a translation. It moves point $\alpha(A)$ into $\alpha\tau\alpha^{-1}\alpha(A) = \alpha(B)$ so it is the translation $\tau_{\alpha(B)-\alpha(A)}$.
3. Let $\rho = \rho_{P,\Theta}$ where $\Theta \neq 0$. (The case $\rho = \iota$ is trivial.) Clearly $\alpha\rho\alpha^{-1}$ is an even isometry with fixed point $\alpha(P)$, so $\alpha\rho\alpha^{-1}$ must be some rotation about point $\alpha(P)$, say $\alpha\rho\alpha^{-1} = \rho_{\alpha(P),\Phi}$. All we need to prove is that $\Phi = \pm\Theta$ where the sign depends on the parity of α.

   Assume first that $\alpha = \sigma_k$ for some line k. Let m and l be lines through point P such that the directed angle from l to m is $\Theta/2$, so $\rho = \sigma_m\sigma_l$. We are free to choose lines m and l in such a way that neither is parallel to k. Let Θ_1 and Θ_2 be the directed angles from m to k and from k to l, respectively. Notice that $\Theta_1 + \Theta_2$ is then the directed angle from m to l, that is, $\Theta_1 + \Theta_2 = -\Theta/2$, at least modulo 180°.



We have

$$\rho_{\alpha(P),\Phi} = \alpha\rho\alpha^{-1} = \sigma_k\sigma_m\sigma_l\sigma_k.$$

This is the product of two rotations $\sigma_k\sigma_m$ and $\sigma_l\sigma_k$ of angles $2\Theta_1$ and $2\Theta_2$, respectively. According to Theorem 2.21 the product is a rotation by $2\Theta_1 + 2\Theta_2 = -\Theta$, that is, $\Phi = -\Theta$ as required.

Assume then a general α. We know that all isometries are products of (at most three) reflections, so $\alpha = \sigma_1\sigma_2\ldots\sigma_n$ for some reflections $\sigma_1, \sigma_2, \ldots, \sigma_n$. Number n is even iff isometry α is even. As

$$\alpha\rho\alpha^{-1} = \sigma_1\sigma_2\ldots\sigma_n\rho\sigma_n\sigma_{n-1}\ldots\sigma_1$$

we can apply the single reflection case n times. In each application the sign of the rotation angle changes, so in the end we have that $\alpha\rho\alpha^{-1}$ is a rotation by the angle $(-1)^n\Theta$.

4. Let $\gamma = \gamma_{m,B-A}$, where $A \neq B$. (If $A = B$ then γ is a reflection, and that was already taken care of.) Then $\alpha\gamma\alpha^{-1}$ is an odd isometry, so it is a glide reflection, say γ'. Because $\gamma'(\alpha(m)) = \alpha(m)$, line $\alpha(m)$ must be the axis of γ'. To find the glide vector of γ' we can make the calculation

$$\gamma'\gamma' = \alpha\gamma\alpha^{-1}\alpha\gamma\alpha^{-1} = \alpha\gamma^2\alpha^{-1} = \alpha\tau_{B-A}\tau_{B-A}\alpha^{-1} = \alpha\tau_{B-A}\alpha^{-1}\alpha\tau_{B-A}\alpha^{-1} = \tau_{\alpha(B)-\alpha(A)}\tau_{\alpha(B)-\alpha(A)},$$

which shows that $\alpha(B) - \alpha(A)$ is the glide vector of γ'.

□

Let $s \subseteq \mathbb{R}^2$. The following terminologies are widely used: If σ_m is a symmetry of s then m is called a line of symmetry for s. If σ_P is a symmetry of s then P is a point of symmetry for s. If $\rho_{C,\Theta}$ is a symmetry of s then C is a center of symmetry and, more precisely, if $\Theta = \frac{360^\circ}{n}$ then C is a center of n-fold symmetry.

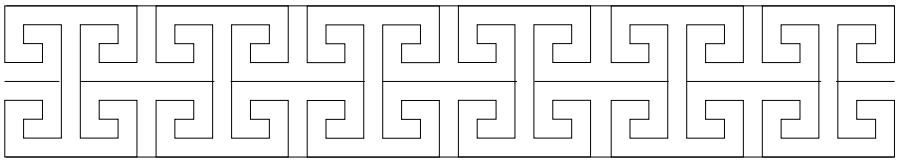
In analyzing symmetries we frequently apply the statements of the conjugacy theorem above in the following forms. Let α be an arbitrary symmetry of set s. Then in s:

- If m is a line of symmetry then also $\alpha(m)$ is a line of symmetry.
- If P is a point of symmetry then also $\alpha(P)$ is a point of symmetry.
- If C is a center of (n-fold) symmetry then also $\alpha(C)$ is a center of (n-fold) symmetry.

2.9 Frieze groups

Let us denote the set of translations by $\mathcal{T}$. It is easily seen to be a subgroup of $\mathcal{I}$. The intersection of two subgroups is also a subgroup, so for every subgroup G of $\mathcal{I}$, the set $G \cap \mathcal{T}$ that contains the translations of G is a subgroup of G, called the translation group of G.

We say that $G \subseteq \mathcal{I}$ is a frieze group if its translation group is cyclic and non-trivial, that is, if the translations are generated by a single translation $\tau \neq \iota$. The name comes from the fact that frieze groups are the symmetry groups of repetitive friezes (=ornamented bands on buildings) such as, for example



(where the pattern is repeated indefinitely in both directions). Notice that there must exists the shortest translation that keeps the frieze invariant — otherwise its symmetry group is not a frieze group. For example, the symmetry group of a horizontal line is not a frieze group as it contains all horizontal translations. It turns out that there are only seven different frieze groups (when we ignore the position, orientation and the size of the frieze) and each is the symmetry group of some $s \subseteq \mathbb{R}^2$.

In this section we make the following convention: The direction of the translations in the frieze group is called the horizontal direction, and the perpendicular direction is then the vertical direction. We start with the following key observation:

Lemma 2.26 *Let G be a subgroup of $\mathcal{I}$ such that all translations in G are horizontal, and assume that there is at least one non-trivial translation. (This includes all frieze groups, but also groups without a shortest translation.) Then there exists a horizontal line m such that all elements of G are products of reflections in vertical lines, possibly followed by the reflection σ_m in line m. These products are:*

- *horizontal translations,*
- *reflections in vertical lines,*
- *reflection σ_m in line m,*
- *halfturns about points of line m, and*
- *glide reflections with axis m.*

Proof. Let $\tau \in G$ be a fixed non-trivial translation.

First, let us prove that all non-trivial rotations in G are halfturns. Let $\rho = \rho_{P,\Theta} \in G$ be arbitrary. Let $A = \tau(P)$, so $A \neq P$, and let $B = \rho(A)$. According to Theorem 2.25, $\rho\tau\rho^{-1}$ is the translation that moves point $\rho(P) = P$ to point $\rho(A) = B$. Translations τ and $\rho\tau\rho^{-1}$ are horizontal, so points A, P and B must be on the same line. This is possible only if ρ is the trivial rotation or the halfturn about P.

Next, let us prove that all reflections in G are in vertical and horizontal lines. Let $\sigma_l \in G$ be arbitrary, P a point of line l, $A = \tau(P)$, and $B = \sigma_l(A)$. According to Theorem 2.25, $\sigma_l\tau\sigma_l^{-1}$ is the translation that moves point $\sigma_l(P) = P$ to point $\sigma_l(A) = B$. Again, translations τ and $\sigma_l\tau\sigma_l^{-1}$ are horizontal, so points A, P and B must be on the same horizontal line. Either $A = B$, in which case A is on line l so l is horizontal, or $A \neq B$, in which case l is the perpendicular bisector of AB so l is vertical.

Finally, let us show that glide reflections of G are horizontal. Indeed, if $\gamma \in G$ is a glide reflection with a non-zero glide A, then γ^2 is the translation with the translation vector $2A$. Vector $2A$ is horizontal, so also the glide A is horizontal.

In the three paragraphs above we have shown that every element of G is a product of reflections in vertical and horizontal lines. As products of reflections in perpendicular directions commute (Corollary 2.12), every element of G is a product

$$\sigma_1\sigma_2\ldots\sigma_v\sigma^{(1)}\sigma^{(2)}\ldots\sigma^{(h)}$$

where each σ_i is a reflection in a vertical line, and each $\sigma^{(j)}$ is a reflection in a horizontal line. Moreover, as the product of three reflections in parallel lines is a reflection in a parallel line, we can reduce the number of reflections so that $v, h \leq 2$.

Next we prove that, in fact, $h \leq 1$. Assume the contrary: some

$$\alpha = \sigma_1\sigma_2\ldots\sigma_v\sigma^{(1)}\sigma^{(2)} \in G$$

where the reflections $\sigma^{(1)}$ and $\sigma^{(2)}$ are in two different horizontal lines. If $v = 1$ then α is a glide reflection with a non-zero vertical glide, and if $v = 0$ or $v = 2$ then α is a translation in a direction that is not horizontal. These isometries do not exist in G, so we must have $h \leq 1$.

Moreover, the possible reflection $\sigma^{(1)}$ in a horizontal line must be in the *same* horizontal line m for all isometries of G. Namely, if G would contain two isometries $\alpha = \alpha'\sigma^{(1)}$ and $\beta = \beta'\sigma^{(2)}$ where α' and β' are products of reflections in vertical lines and $\sigma^{(1)}$ and $\sigma^{(2)}$ are reflections in two different horizontal lines then the product

$$\alpha\beta = \alpha'\sigma^{(1)}\beta'\sigma^{(2)} = \alpha'\beta'\sigma^{(1)}\sigma^{(2)}$$

would contradict the previous paragraph.

So we conclude that every element of G is a product of 0,1 or 2 reflections in vertical lines, or a product of 0,1 or 2 reflections in vertical lines followed by σ_m, the reflection in the horizontal axis m of the group. This leaves the following non-trivial possibilities:

- σ_m: the reflection in the axis m,
- $\sigma_1\sigma_m$: a halfturn about a point of line m,
- $\sigma_1\sigma_2\sigma_m$: a glide reflection with axis m,
- σ_1: a reflection in a vertical line, and
- $\sigma_1\sigma_2$: a horizontal translation

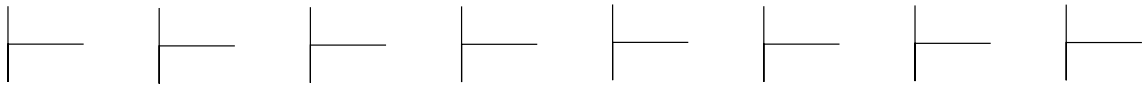
□

Now we are ready to classify all frieze groups. Let G be a frieze group whose translations are generated by the shortest translation τ_A, and let m be the horizontal line from the previous lemma, called the axis of the frieze group. Let $2d$ be the length of vector A, so that τ_A is a product of two reflections in vertical lines at distance d. The translations in G are then exactly the products of two reflections in any two vertical lines whose distance is a multiple of d.

1) Assume first that $\sigma_m \in G$. Let l and k be arbitrary vertical lines. Then

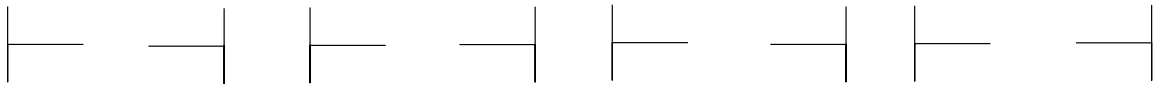
$$\begin{aligned} \sigma_l\sigma_m \in G &\iff \sigma_l \in G, \text{ and} \\ \sigma_l\sigma_k\sigma_m \in G &\iff \sigma_l\sigma_k \in G, \end{aligned}$$

so glide reflections of G are uniquely determined by the translations, and the reflections in vertical lines are uniquely determined by the halfturns. If there are no halfturns in G then G is generated by τ_A and σ_m, and it is the symmetry group of the infinite strip



Let us call it the group F_{1001}. The four indices 1001 are interpreted as follows: the group contains the reflection σ_m in the axis m, does not contain any reflection σ_l in vertical lines, does not contain any halfturns, and contains glide reflections. This information uniquely specifies the group (for given A and m).

Assume then that there is some halfturn σ_P in G. We know that point P is on line m. Because $\sigma_P\sigma_Q = \tau_{2(P-Q)}$ is a translation, the other halfturns are now uniquely determined: they are at points of line m whose distance from P is a multiple of d. These then also uniquely determine the reflections σ_l in vertical lines. Group $G = F_{1111}$ is the symmetry group of

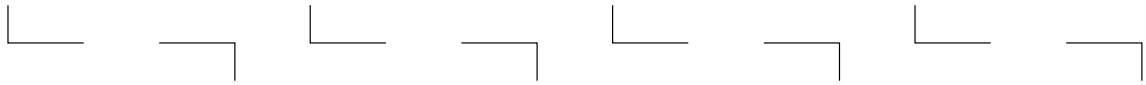


Groups F_{1001} and F_{1111} are the only groups containing the reflection σ_m.

2) Consider then groups that do not contain σ_m. One alternative is that there are no isometries except the translations: We have the symmetry group F_{0000} of the strip

Let us assume then that there are other symmetries. The product of a reflection in a vertical line and a halfturn is a glide reflection, the product of a glide reflection and a halfturn is a reflection in a vertical line, and the product of a glide reflection and a reflection in a vertical line is a halfturn. Conclusion: G either contains all three types of isometries, or at most one of the types. There are four alternatives, resulting in groups F_{0100}, F_{0010}, F_{0001} and F_{0111}, as discussed below.

If G contains a halfturn $\sigma_P = \sigma_l\sigma_m$ then the other halfturns are uniquely determined: they are the products of σ_P and the translations in G. The distances between the centers of the halfturns are then exactly the multiples of d. This means that group F_{0010} is uniquely determined, and it is the symmetry group of the following strip:



Analogously, if G contains a reflection σ_l in a vertical line l then the other reflections are determined as they must be the products of σ_l and the translations in G. The lines of the reflections are at distances that are multiples of d. So we have the group F_{0100} which is the symmetry group of

Consider then a glide reflection $\gamma = \sigma_l\sigma_k\sigma_m \in G$ with axis m. Let $2g$ be the length of its glide vector, that is, g is the distance between lines l and k. Then g must be a multiple of $d/2$ as γ^2 is a translation of length $4g$. On the other hand, g cannot be a multiple of d because then there would exist a translation in G that would cancel the glide, leaving σ_m, and we assumed that σ_m is not in G. We conclude that g must be an odd multiple of $d/2$, or equivalently, the length $2g$ of the glide is an odd multiple of d. All such glide reflections are obtained from γ by multiplying it with translations, so we have completely characterized the glide reflections. Group F_{0001} is the symmetry group of

The last open possibility is that G contains halfturns, reflections in vertical lines and glide reflections. As discussed above, the glide reflections are uniquely determined (the glides are by odd multiples of d), and after we fix one center P of a halfturn, also the halfturns are uniquely determined. This also fixes the reflections as they are the products of the glide reflections and the halfturns. The lines of the reflections bisect the consecutive points of reflections. We have the group F_{0111}, which is the symmetry group of the following strip:

We have fully classified the frieze groups, and we found seven different types. In each case, a "frieze" with the given symmetries was given, to prove that the seven types of frieze groups are the symmetry groups of some sets $s \subseteq \mathbb{R}^2$. Notice that each of the seven groups has infinitely many "geometric realizations", as the axis m can be any line, the shortest translation τ can be any non-trivial translation parallel to m, and in those groups that involve halfturns or reflections in vertical line, one center P of a halfturn or one line l of a reflection can be selected. But modulo these parameters, the groups are unique. It is clear that all realizations of each group are isomorphic, and even more than that, isomorphic by isomorphisms that preserve the type of isometry (translations correspond to translations, reflections to reflections, rotations to rotations,...).

We have proved the following theorem:

Theorem 2.27 *Let G be a frieze group whose translations are generated by τ. Then there exists a line m parallel to τ, and if G contains a halfturn there exists a point $P \in m$, otherwise a line l perpendicular to m, such that G is one of the following seven groups:*

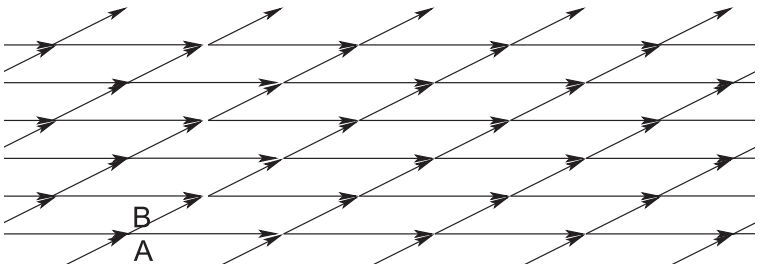
$$\begin{array}{lllllllll} F_{0000} & = & \langle\tau\rangle & & F_{1001} & = & \langle\tau,\sigma_m\rangle & & F_{1111} & = & \langle\tau,\sigma_m,\sigma_P\rangle \\ F_{0100} & = & \langle\tau,\sigma_l\rangle & & F_{0010} & = & \langle\tau,\sigma_P\rangle & & & & \\ F_{0001} & = & \langle\gamma\rangle & & F_{0111} & = & \langle\gamma,\sigma_P\rangle & & & & \end{array}$$

where γ is the glide reflection with axis m such that $\gamma^2 = \tau$.

□

2.10 Wallpaper groups

A wallpaper group G is a subgroup of $\mathcal{I}$ whose translations are generated by two non-parallel translations τ_1 and τ_2. Translations commute with each other, so the translations of G are exactly the isometries $\tau_1^i\tau_2^j$ for all integers i and j. If A and B are the vectors of translations τ_1 and τ_2 then the vectors of translations $\tau_1^i\tau_2^j$ are $iA + jB$, which form a lattice



Let us first show that there exists a shortest translation in G.

Lemma 2.28 *Wallpaper group G has a shortest non-trivial translation. More generally, any non-empty subset s of translations of G contains a shortest non-trivial translation.*

Proof. Let A and B be the translation vectors of the generating translations τ_1 and τ_2. Let

$$B = rA + B' \qquad \text{and} \qquad A = qB + A'$$

be the decompositions of vectors A and B into a sum of orthogonal vectors, where $r, q \in \mathbb{R}$ and $B' \perp A$ and $A' \perp B$. As A and B are not parallel, vectors A' and B' are non-zero. Let $a > 0$ and $b > 0$ be the lengths of vectors A' and B', respectively.

Consider an arbitrary translation vector $A_{ij} = iA + jB$ in G. Using the orthogonal decompositions above we have

$$A_{ij} = (i + jr)A + jB' \qquad \text{and} \qquad A_{ij} = (j + iq)B + iA'.$$

These are sums of two orthogonal vectors, so the length of A_{ij} is at least $|j|b$, the length of jB', and at least $|i|a$, the length of iA'. Let c be the length of some vector X in the set s of translations we consider. Then any vector A_{ij} with $|j| > c/b$ or $|i| > c/a$ is longer than vector X. Therefore there are only a finite number of vectors that can potentially be shorter than X. The shortest among them is the shortest translation vector in set s. □

Rosette groups, frieze groups and wallpaper groups are exactly the discrete symmetry groups: We call a subgroup G of $\mathcal{I}$ discrete if it does not contain arbitrarily short translations and does not contain arbitrarily small rotations. More precisely, G is discrete if there exists $\varepsilon > 0$ such that

$$\begin{aligned} 0 < |A| < \varepsilon \quad &\Longrightarrow \quad \tau_A \notin G, \text{ and} \\ 0 < \Theta < \varepsilon \quad &\Longrightarrow \quad \rho_{C,\Theta} \notin G. \end{aligned}$$

($|A|$ is the length of the translation vector A.)

Theorem 2.29 *Discrete subgroups of $\mathcal{I}$ are exactly the rosette groups, frieze groups and wallpaper groups.*

Proof. ($\Longleftarrow$) Rosette groups are finite and hence discrete. In frieze groups, the translation that generates all translations is the shortest one, and halfturns are the only possible rotations, so frieze groups are discrete. Let G be a wallpaper group. By Lemma 2.28, it contains a shortest translation τ. For every rotation $\rho \in G$, the isometry $\tau' = \rho\tau\rho^{-1}$ is the translation that maps the center C of ρ to $\rho\tau(C)$. Consequently, translation $\tau'\tau^{-1}$ takes point $\tau(C)$ into $\rho\tau(C)$. This translation is arbitrarily short for arbitrarily small rotation angles, so G cannot contain arbitrarily small rotations. Hence G is discrete.

($\Longrightarrow$) Let G be a discrete subgroup of $\mathcal{I}$.

(1) If G contains no non-trivial translations then it does not contain any glide reflections with non-zero glide vector. There are only rotations and reflections in G. Rotations can only have a finite number of different rotation angles as otherwise there would be arbitrarily small rotations in G. Two rotations by the same angles but with different centers generate a translation, so the rotations of G have the same center C. Reflection lines must contain C, and there are only a finite number of possible angles between the lines of reflections. We conclude that the group is finite, and hence it is a rosette group.

(2) Suppose then that G contains a non-trivial translation τ_A. Due to discreteness there can be only a finite number of different translations by vectors shorter than A, so a shortest non-trivial translation exists. We may assume τ_A is a shortest translation.

(2a) If all translations in G are generated by τ_A then G is a frieze group.

(2b) Suppose then that there exists a translation τ_B in G that is not generated by τ_A. Again, by discreteness, a shortest such translation exists, so we may assume that B has minimum length. To complete the proof of Theorem 2.29, we use the following lemma that states that τ_A and τ_B generate all translations in G, implying that G is a wallpaper group.

Lemma 2.30 *Let G be a discrete subgroup of $\mathcal{I}$, let τ_A be a shortest non-zero translation in G, and let $\tau_B \in G$ be a shortest translation not generated by τ_A. Then τ_A and τ_B generate all translations of G.*

Proof. It is clear that vectors A and B are not in parallel directions (otherwise A would not be the shortest translation vector), so every vector of $\mathbb{R}^2$ is a linear combination of A and B. Assume that group G contains a translation τ_C such that $\tau_C \notin \langle \tau_A, \tau_B \rangle$. Let $C = xA + yB$ be the representation of C as a linear combination of vectors A and B, where $x, y \in \mathbb{R}$. By subtracting integer multiples of vectors A and B from vector C, we can reduce x and y so that $-\frac{1}{2} \leq x, y \leq \frac{1}{2}$. But then, using the triangular inequality, we obtain

$$|C| = |xA + yB| \leq |x||A| + |y||B| \leq (|A| + |B|)/2 \leq |B|.$$

The first inequality can be an equality only if $x = 0$ or $y = 0$, but in these cases the second inequality is proper. So in each case: $|C| < |B|$, which contradicts the minimality of vector B. □

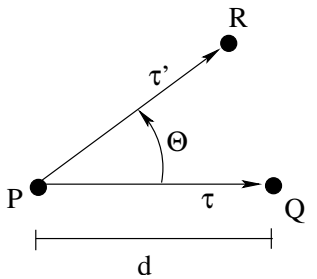
Let us start analyzing the possibilities for the wallpaper groups. It turns out that there are 17 different types of groups. Deriving them is a lengthy case analysis. The rest of this chapter provides a complete derivation. (See also the slide presentation from the course web page.)

Our first observation is an important restriction on possible rotations in wallpaper groups:

Theorem 2.31 (Crystallographic restriction) *A wallpaper group G can only contain rotations by multiples of $60°$ and $90°$. Hence all centers of rotations are centers of n-fold rotations for $n = 2, 3, 4$ or 6. Moreover, a 4-fold rotation cannot co-exist with 3- or 6-fold rotations.*

Proof. Let $\tau = \tau_A$ be the shortest translation in G, and let d be the length of its translation vector A.

Let $\rho = \rho_{P,\Theta} \in G$ be a non-trivial rotation, and let $Q = \tau(P)$ and $R = \rho(Q)$. Then G contains also the translation $\tau' = \rho\tau\rho^{-1}$ that moves point $\rho(P) = P$ to point $\rho(Q) = R$. Translation $\tau'\tau^{-1}$ then moves point Q to point R.



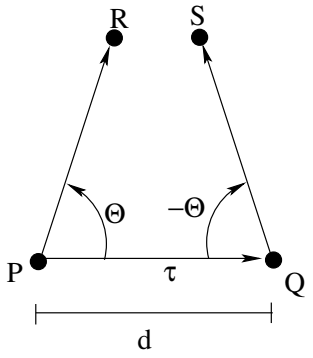
If $0° < \Theta < 60°$ then the distance between points Q and R is less than d, which contradicts the fact that τ is the shortest translation in G. We conclude that every non-trivial rotation is by an angle that is at least $60°$. This also means that G can contain at most 6 different rotations about point P, because if we would have rotations by angles $\Theta_1, \Theta_2, \ldots, \Theta_7$ where

$$0° \leq \Theta_1 < \Theta_2 < \cdots < \Theta_7 < 360°$$

then necessarily $0^\circ < \Theta_{i+1} - \Theta_i < 60^\circ$ for some $i = 1, 2, \ldots 6$, a contradiction.

Let Θ be the smallest positive rotation angle about point P, and let Φ be any other rotation angle about P. There exists an integer k such that $0 \leq \Phi - k\Theta < \Theta$. This implies that $\Phi = k\Theta$. Therefore the rotations about point P are generated by $\rho_{P,\Theta}$, and $\Theta = \frac{360^\circ}{n}$ for some $n \leq 6$.

We still have to show that the case $n = 5$ of five-fold rotations is not possible. The rotation angle of a five-fold rotation is $\Theta = 72^\circ$. Consider points P, Q and R as in the beginning of the proof. Point Q is the center of rotation $\tau\rho\tau^{-1}$ by the same angle Θ, and therefore G contains the rotation of $-\Theta$ about Q. Let $S = \rho'(P)$.



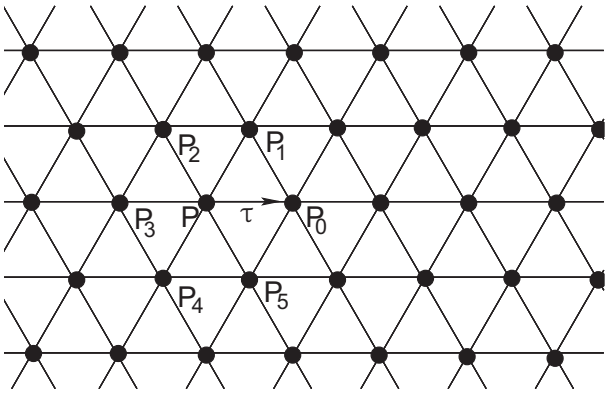
It is easily seen that the distance between points R and S is positive but less than d for angles in the interval $60^\circ < \Theta < 90^\circ$. In particular, this includes the case $\Theta = 72^\circ$ of five-fold rotations. Since G must contain the translation that moves R to S, this contradicts the minimality of distance d.

Finally we easily observe that if G contains a rotation ρ of 90° then it cannot contain any rotation ρ' of 60° or 120° because $\rho^{-1}\rho'$ would be a rotation whose angle is $\pm 30^\circ$.

□

Let us start analyzing different wallpaper groups case-by-case depending on the largest order of rotation that G contains.

1) Assume that G contains a 6-fold rotation $\rho = \rho_{P,60^\circ}$. Let τ be the shortest translation in G, let d be its length, and let $P_0 = \tau(P)$. Rotating point P_0 about point P defines points $P_i = \rho^i(P_0)$ for $i = 1, 2, \ldots 5$ such that all translations $\tau_{P_i - P}$ are in G. Then each P_i is a center of a 6-fold rotation in G. These isometries are all generated by ρ and τ through conjugacies. We can repeat the reasoning on all P_i, and then again on the six centers of rotation around them and so on. We conclude that G contains 6-fold rotations about centers that are the vertices of a lattice of equilateral triangles, and G contains all translations between vertices of the lattice. Let us denote by s_6 the set of the lattice points, indicated by black circles in the following figure:

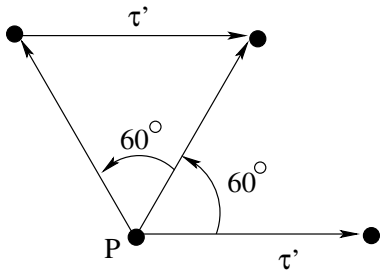


Let us show that the even isometries in G are exactly the even symmetries of s_6. First, there can be no translation that moves a lattice point into a non-lattice point: The distance from every point of the plane to the closest lattice point is less than d, so if τ' is a translation that moves lattice point P into a non-lattice point $Q = \tau'(P)$ then the translation that moves Q to its closest lattice point is in G and it is shorter than τ, which contradicts the minimality of τ. So the translations of G are exactly that translations that keep s_6 invariant.

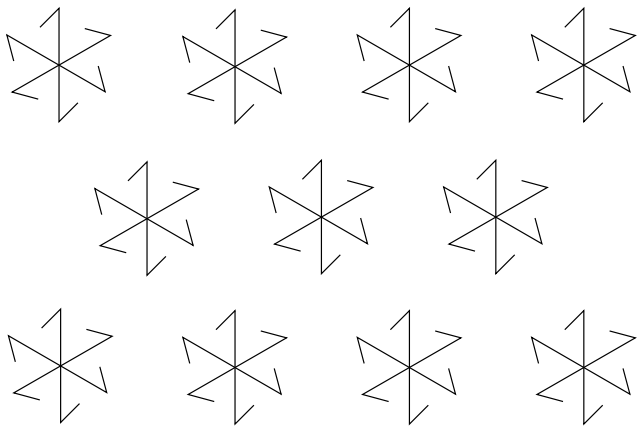
Consider then an arbitrary rotation $\rho' \in G$. The crystallographic restriction states that ρ' is a 2-, 3- or 6-fold rotation. This means that $\rho'\rho^i$ is a translation for some integer i. Since translations in G are symmetries of s_6, and since ρ is a symmetry of s_6 we conclude that ρ' is also a symmetry of s_6.

Conversely, if ρ' is any rotation in the symmetry group of s_6 then it must be a 2-, 3- or 6-fold rotation (as the symmetry group of s_6 is a wallpaper group that contains 6-fold rotations) so $\rho'\rho^i$ is a translation for some integer i. As ρ is a symmetry of s_6 this translation is also a symmetry of s_6. All such translations are in G, so $\rho' \in G$ as well.

We have proved that the even elements of G are exactly the even symmetries of s_6. If there are no odd isometries in G we have our first wallpaper group $W_6 = \langle \tau, \rho_{P,60^\circ} \rangle$ that consists of the even symmetries of s_6. In addition to the translations and 6-fold rotation about lattice points this group also contains 3-fold rotations about the centers of the equilateral triangles, and 2-fold rotations about the midpoints between adjacent lattice points. Notice that the lattice points are the only centers of 6-fold rotations, because if ρ' is a 60° rotation then $\rho'\rho^{-1} = \tau'$ is a translation and, since $\tau'\rho\rho\tau'(P) = \rho\tau'(P)$, the lattice point $\rho\tau'(P)$ is the fixed point of $\rho' = \tau'\rho$.

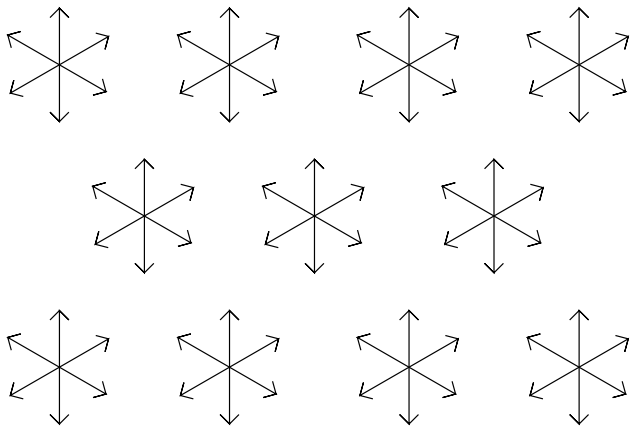


Group W_6 is the symmetry group of the following pattern where odd isometries are prevented by "directing" the lattice points counter-clockwise:

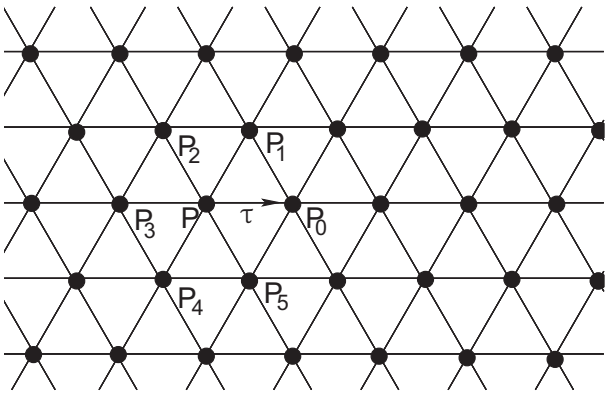


Assume then that G also contains some odd isometry α. This isometry has to take 6-fold rotation centers of G into 6-fold rotation centers of G, that is, α is a symmetry of s_6. If β is any other odd symmetry of s_6 then $\alpha\beta \in G$ as $\alpha\beta$ is an even symmetry of s_6, so also $\beta \in G$. Conclusion: G is the symmetry group of

s_6. Note that s_6 has odd symmetries (e.g. a reflection σ in any line through two closest lattice points), so we have a new wallpaper group $W_6^1 = \langle \tau, \rho_{P,60^\circ}, \sigma \rangle$. Set s_6 is an example of a pattern whose symmetry group is W_6^1. Here is another one:

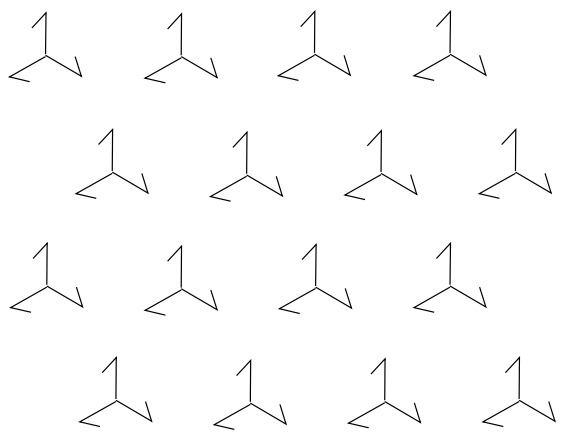


2) Assume that G contains a 3-fold rotation $\rho = \rho_{P,120^\circ}$ but no 6-fold rotations. We start in the same way as with the 6-fold rotations: Let τ be the shortest translation in G, let d be its length, and let $P_0, P_1, \ldots, P_5$ be the points where P is taken by the translations τ, $\rho^{-1}\tau^{-1}\rho$, $\rho\tau\rho^{-1}$, τ^{-1}, $\rho^{-1}\tau\rho$ and $\rho\tau^{-1}\rho^{-1}$, respectively. Points $P_0, P_1, \ldots, P_5$ are the vertices of the regular hexagon with center P, and they are all centers of 3-fold rotations in G. We can repeat the reasoning on each P_i instead of P, so we obtaining again a lattice of equilateral triangles such that the vertices of the lattice are centers of 3-fold rotations, and the translations that move lattice points to lattice points are in G.



As before, let s_6 be the set of vertices of this lattice. Next we show that the even isometries of G are exactly those symmetries of s_6 that are translations or 3-fold rotations. First, exactly as in the case of W_6, we see that no other translation is possible: a translation that moves a lattice point into a non-lattice point contradicts the minimality of translation τ. So the translations of G are exactly the translations that keep s_6 invariant. Consider then a rotation in G. We assumed that there are no 6-fold rotations, and therefore there can be no 2-fold rotations either (together with a 3-fold rotation any 2-fold rotation generates a 6-fold rotation). All other rotations would contradict the crystallographic restriction, so all rotations in G are 3-fold. Conversely, every 3-fold rotation ρ' that keeps s_6 invariant must be in G because $\rho'\rho^{-1}$ is a translation that keeps s_6 invariant, and all such translations are in G.

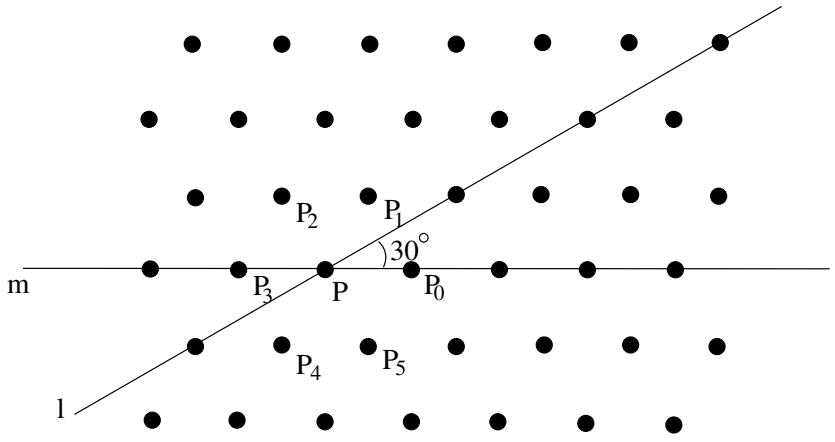
If there are no odd isometries in G we have our third wallpaper group $W_3 = \langle \tau, \rho_{P,120^\circ} \rangle$. In addition to the translations and 3-fold rotations about lattice points, group W_3 also contains 3-fold rotations about the centers of the equilateral triangles of the lattice. Group W_3 is the symmetry group of the following pattern:



Assume then that G also contains odd isometries. If G contains a glide reflection γ then it also contains a reflection because $\gamma\rho\gamma\rho^{-1}\gamma$ is a reflection for every glide reflection γ and 3-fold rotation ρ (homework). Every line p of reflection must contain a center of 3-fold rotation because also $\rho(p)$ is a line of reflection, lines p and $\rho(p)$ are not parallel so they intersect, and the product $\sigma_p\sigma_{\rho(p)}$ is a rotation about the point of intersection. In the beginning of case 3 the first center P of the 3-fold rotation ρ was chosen arbitrarily, so we may assume that P is on line p. Consequently P is a fixed point of a reflection in G.

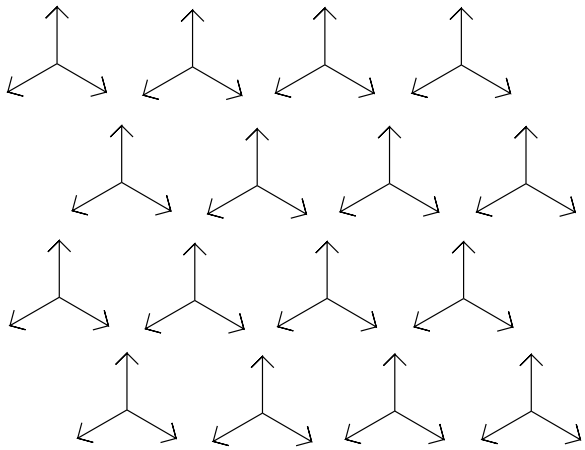
It follows then that every odd isometry in G is a symmetry of s_6. Assume the contrary: there is an odd $\alpha \in G$ and a lattice point Q such that $\alpha(Q)$ is not a lattice point. Then $\alpha\tau_{Q-P}\sigma_p \in G$ is an even isometry that moves point P into the non-lattice point $\alpha(Q)$, and this contradicts the fact that all even isometries in G are symmetries of s_6.

Let m be a line through two adjacent lattice points P and P_0, and let l be the line through P such that the angle from m to l is 30°.

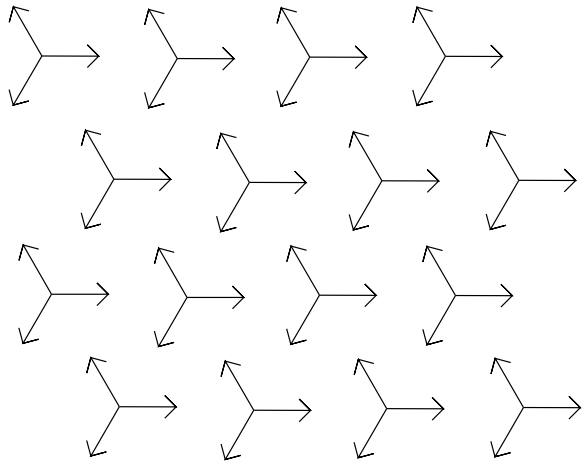


Both σ_m and σ_l are symmetries of s_6, but since $\sigma_l\sigma_m$ is a rotation by 60° they cannot both be in group G. Let us prove that G must contain one of them. Assume the contrary: neither σ_m nor σ_l is in G, and let α be some odd isometry in G. Then $\alpha\sigma_m$ and $\sigma_l\alpha^{-1}$ are even symmetries of s_6 that do not belong to G, so they have to be rotations by an angle that is an odd multiple of 60° (=by 60, 180 or -60 degrees). Their product $\sigma_l\alpha^{-1}\alpha\sigma_m = \sigma_l\sigma_m$ would then be a translation or a rotation by an even multiple of 60°, but we know that $\sigma_l\sigma_m$ is a rotation by 60°, a contradiction. We conclude that exactly one of the reflections σ_m and σ_l is in G.

Once we know one odd element of G, all other odd elements are uniquely determined by the even elements of G. We have two new wallpaper groups: $W_3^1 = \langle \tau, \rho_{P,120^\circ}, \sigma_l \rangle$, which is the symmetry group of



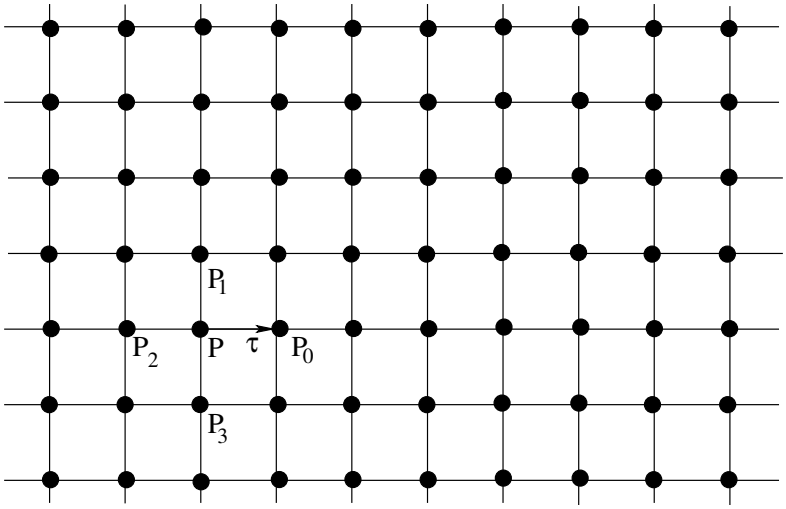
and $W_3^2 = \langle \tau, \rho_{P,120^\circ}, \sigma_m \rangle$, which is the symmetry group of



A difference between these groups is that W_3^1 contains a line of reflection through every center of 3-fold rotation, while in W_3^2 there are lines of symmetry only through some of the rotation centers, namely those that are the lattice points.

3) Let us assume now that G contains a 4-fold rotation $\rho_{P,90^\circ}$. Then it cannot contain 3- or 6-fold rotations. As in the previous cases: let τ be the shortest translation in G, let d be its length, and let P_i be the point where P is taken by the translation $\rho^i \tau \rho^{-i}$, for $i = 0, 1, 2$ and 3. Points P_0, P_1, P_2 and P_3 are all centers of 4-fold rotations in G, so we can repeat the reasoning on each P_i. We obtain an infinite lattice of centers of 4-fold rotations, but this time the lattice is a square lattice instead of a triangular one. (See the next figure.) All translations between lattice points are in group G.

If G would contain any other translations, then it would contain a translation that moves a non-lattice point into the closest lattice point. This is not possible as the distance of every point of the plane from the lattice is less than d, the length of the shortest translation. We conclude that the translations in G are exactly the translations that keep the lattice invariant. Let us denote the points of the square lattice by s_4.

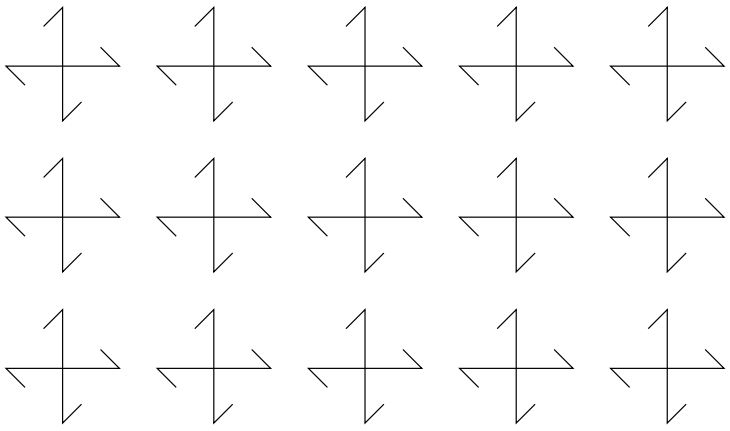


Analogously to the case of 60° rotations, we can prove that the even isometries in G are exactly the even symmetries of s_4. We already know this for translations. Consider then an arbitrary rotation $\rho' \in G$. The crystallographic restriction states that ρ' is a 2- or 4-fold rotation. This means that $\rho'\rho^i$ is a translation for some integer i. Since translations in G are symmetries of s_4, and since ρ is a symmetry of s_4 we conclude that ρ' is also a symmetry of s_4.

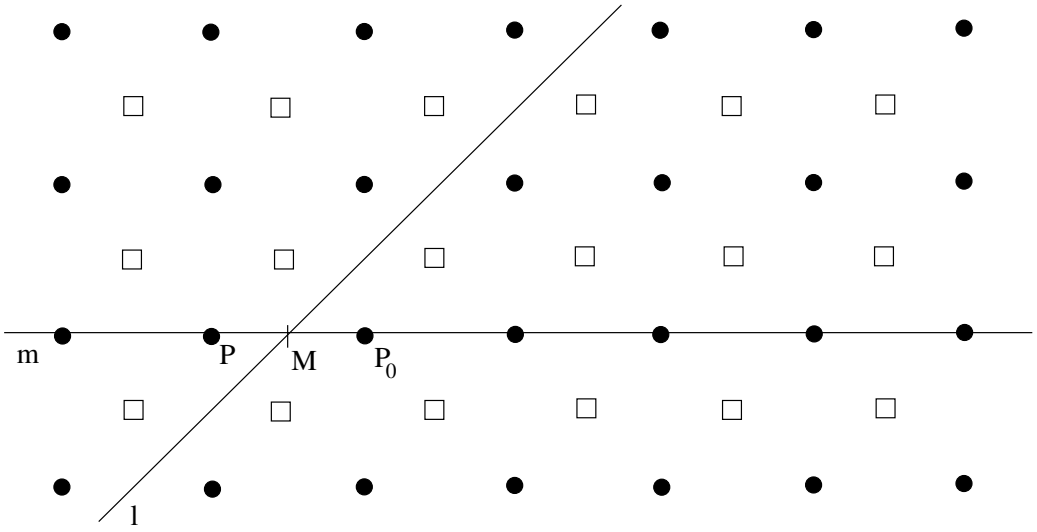
Conversely, if ρ' is any rotation in the symmetry group of s_4 then it must be a 2- or 4-fold rotation. This follows from the crystallographic restriction and the fact that the symmetry group of s_4 is a wallpaper group that contains 4-fold rotations. So $\rho'\rho^i$ is a translation for some integer i and, as ρ is a symmetry of s_4, this translation is also a symmetry of s_4. All such translations are in G, so $\rho' \in G$ as well.

If G contains no odd isometries then G is the group of even symmetries of s_4. This is a new wallpaper group $W_4 = \langle \tau, \rho_{P,90^\circ} \rangle$. In addition to the translations and 4-fold rotations about lattice points this group also contains 4-fold rotations about the centers of the lattice squares, and 2-fold rotations about the midpoints between adjacent lattice points. Let us prove that no other rotations exist in G. Consider a center Q of a halfturn. Lattice point P is also a center of a halfturn. The product of the two halfturns is the translation by vector $2(Q - P)$. Translations are between lattice points, so Q must be a midpoint between lattice points. The only such points are the centers of the lattice squares (which are easily seen to be also centers of 4-fold rotations), and the midpoints between adjacent lattice points (which are easily seen not to be centers of 4-fold rotations). No other rotations are possible.

Group W_4 is the symmetry group of the following pattern:

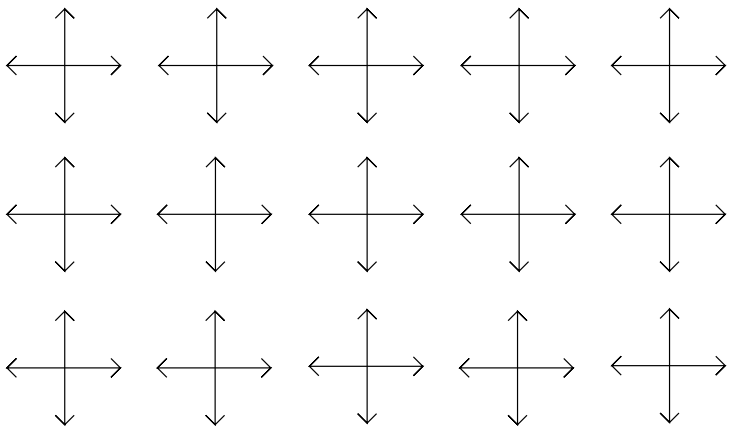


Assume then that G also contains odd isometries. Let m be a line through some adjacent lattice points, and let l be a line that intersects m at 45° in some midpoint M between adjacent lattice points:

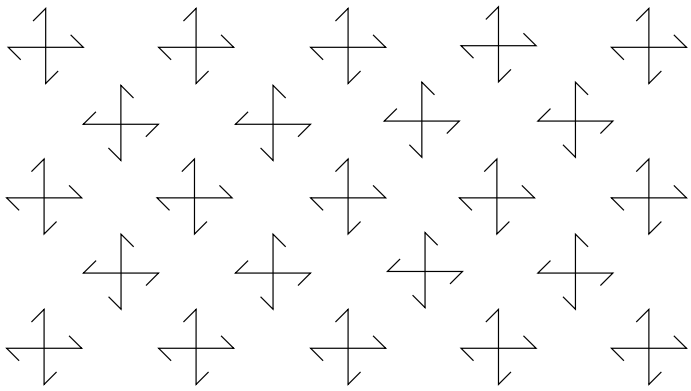


Reflection σ_m is a symmetry of s_4 whereas reflection σ_l is not. Instead, σ_l exchanges lattice points and the centers of the lattice squares. Group G cannot contain both σ_m and σ_l because then it would also contain a 4-fold rotation about point M. Let us prove that G must contain either σ_m or σ_l. If there is an odd isometry $\alpha \in G$ that takes some lattice point into a lattice point then every odd isometry of G must be a symmetry of s_4. (Otherwise there would be an even element in G that is not a symmetry of s_4.) As G contains all even symmetries of s_4 then all odd symmetries of s_4 are in G as well, and this includes σ_m. If, on the other hand, G contains an odd isometry α that takes all lattice points into non-lattice points then these non-lattice points must be the centers of the lattice squares, so $\sigma_l\alpha$ is an even symmetry of s_4. Therefore $\sigma_l\alpha \in G$, and also $\sigma_l \in G$.

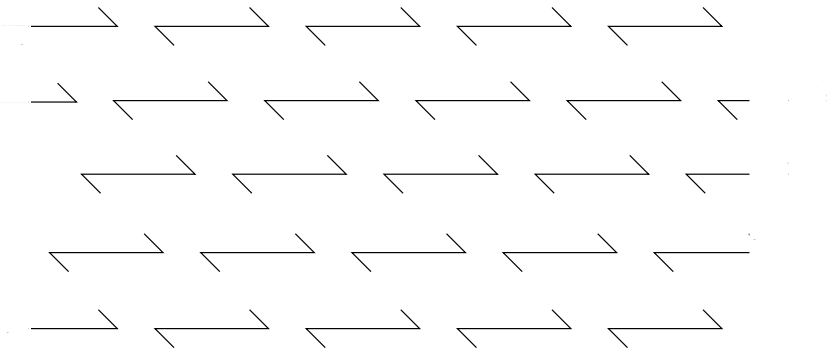
We have two new wallpaper groups $W_4^1 = \langle \tau, \rho_{P,90^\circ}, \sigma_m \rangle$, which is the symmetry group of



and $W_4^2 = \langle \tau, \rho_{P,90^\circ}, \sigma_l \rangle$, which is the symmetry group of



4) Assume that G contains halfturn σ_P, and that all non-trivial rotations in G are halfturns. Let τ_1 and τ_2 be two translations that generate all translations of G. Let the lattice points be the points $\tau_1^i\tau_2^j(P)$ for all integers i, j. They are all centers of halfturns. Also the products of σ_P and the translations $\tau_1^i\tau_2^j$ are halfturns about points that are midpoints between lattice points, that is, centers of the lattice parallelograms as well as the midpoints of their sides. No other halfturns are possible as otherwise we would get translations that are not invariants of the lattice. We conclude that we have found all even isometries in G. If G contains no odd isometries then we have the wallpaper group $W_2 = \langle \tau_1, \tau_2, \sigma_P \rangle$. It is the symmetry group of



Assume then that G contains also some odd isometries. As in the previous cases, a single odd isometry $\alpha \in G$ uniquely determines all odd isometries because they are obtained by multiplying α with the even elements of G. The purpose of the following lemma is to limit the possible odd isometries that any wallpaper group can contain. It turns out that if G contains odd isometries then the translation lattice is rhombic or rectangular:

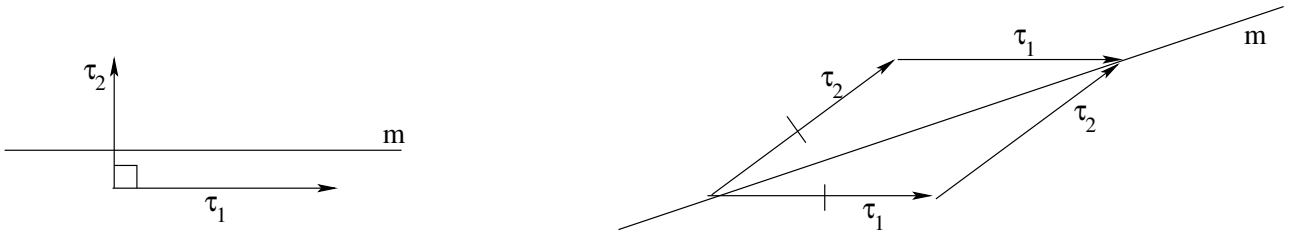
Lemma 2.32 *Let G be a wallpaper group that contains an odd isometry with axis m. Then there exist translations $\tau_1, \tau_2 \in G$ that generate all translations of G and either*

(1) τ_1 is parallel to m and τ_2 is perpendicular to m, or

(2) τ_1 and τ_2 are of equal length and m is parallel to $\tau_1\tau_2$.

Moreover, in case (2), group G contains a reflection.

In case (1) the translation lattice is rectangular, and m is parallel to a side of the rectangles, and in case (2) the translation lattice is rhombic, and m is parallel to a diagonal of the rhombi:



Case (1) Case (2)

Proof. Let τ_A be the shortest translation in G, and let $\tau_B \in G$ be the shortest translation not generated by τ_A. According to Lemma 2.30, τ_A and τ_B generate all translations of G. Let $\alpha \in G$ be an odd isometry with axis m, that is, α is a glide reflection with axis m. Notice that for every translation τ

$$\alpha\tau\alpha^{-1} = \sigma_m\tau\sigma_m.$$

This follows from the facts that $\alpha = \sigma_m\tau'$ where τ' is a translation, and that translations commute.

Consider the translation $\tau_C = \alpha\tau_A\alpha^{-1} = \sigma_m\tau_A\sigma_m$. It has the same length as the shortest translation τ_A. If τ_C is not generated by τ_A then it is the shortest translation not generated by τ_A, and according to Lemma 2.30 translations τ_A and τ_C generate all translations of G. If we choose $\tau_1 = \tau_A$ and $\tau_2 = \tau_C$ we have generating translations that satisfy the condition (2) of the lemma.

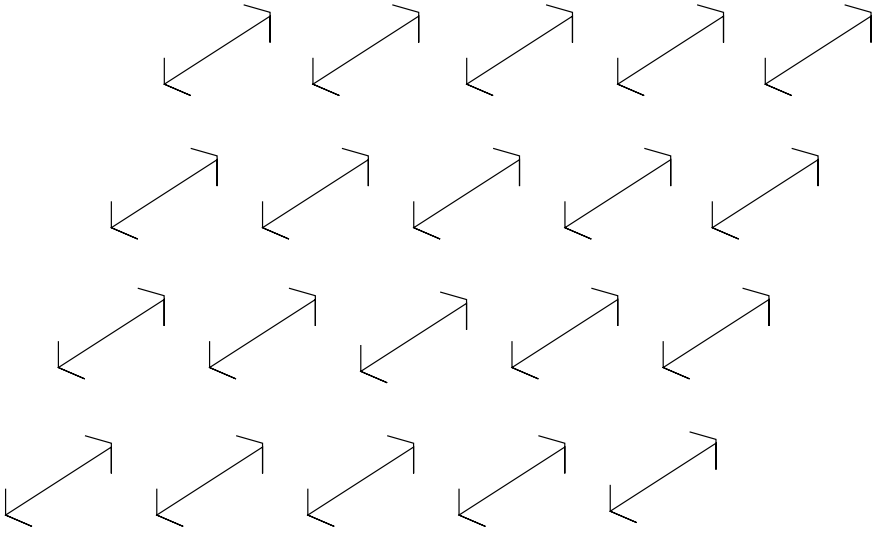
Assume then that τ_C is generated by τ_A. Then either $C = A$, in which case m is parallel to A, or $C = -A$, in which case m is perpendicular to A. Consider the conjugate $\tau_D = \alpha\tau_B\alpha^{-1}$ (if m is parallel to A) or $\tau_D = \alpha\tau_{-B}\alpha^{-1}$ (if m is perpendicular to A). In either case, $B+D$ is parallel to A. If $|B+D| > |A|$ then $B - A$ or $B + A$ is shorter than B, which contradicts the minimality of vector B. We must have $B + D = 0$ or $B + D = \pm A$. If $B + D = 0$ then B is perpendicular to A and we can choose $\tau_1 = \tau_A$, $\tau_2 = \tau_B$ and condition (1) of the lemma is satisfied. And if $B + D = \pm A$ then we choose $\tau_1 = \tau_B$, $\tau_2 = \tau_D$ (if m is parallel to A) or $\tau_1 = \tau_{-B}$, $\tau_2 = \tau_D$ (if m is perpendicular to A). In either case, condition (2) of the lemma is satisfied. Notice that τ_B and τ_D generate all translations because they generate τ_A.

Finally, to prove the last claim, assume that case (2) applies. Because α^2 is a translation that is parallel to $\tau_1\tau_2$, we must have that $\alpha^2 = (\tau_1\tau_2)^i = \tau_2^i\tau_1^i$ for some integer i. Since translations τ_1 and τ_2 are conjugate by $\tau_2 = \alpha\tau_1\alpha^{-1}$, we also have that $\tau_2^i = \alpha\tau_1^i\alpha^{-1}$. This means that $\alpha^2 = \alpha\tau_1^i\alpha^{-1}\tau_1^i$. Divide both sides by α^2 from the left, and we have the result that $\alpha^{-1}\tau_1^i$ is an odd involution, that is, a reflection.

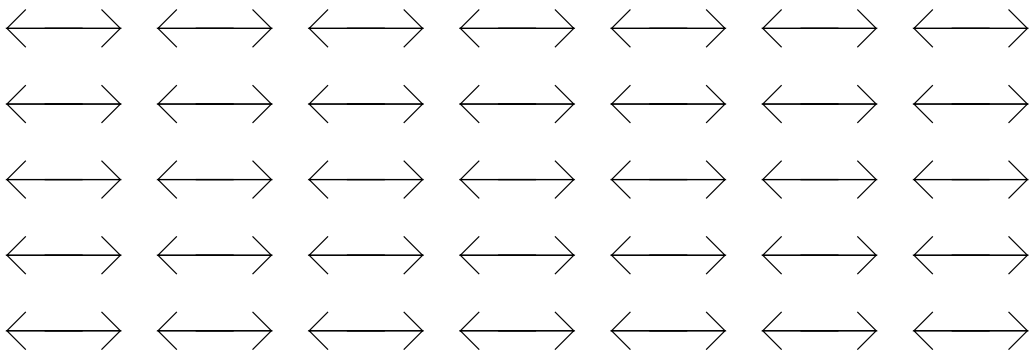
□

Our lemma limits the number of possible odd isometries of wallpaper groups sufficiently so that we can proceed with the analysis of the wallpaper groups G with halfturns and some odd isometries.

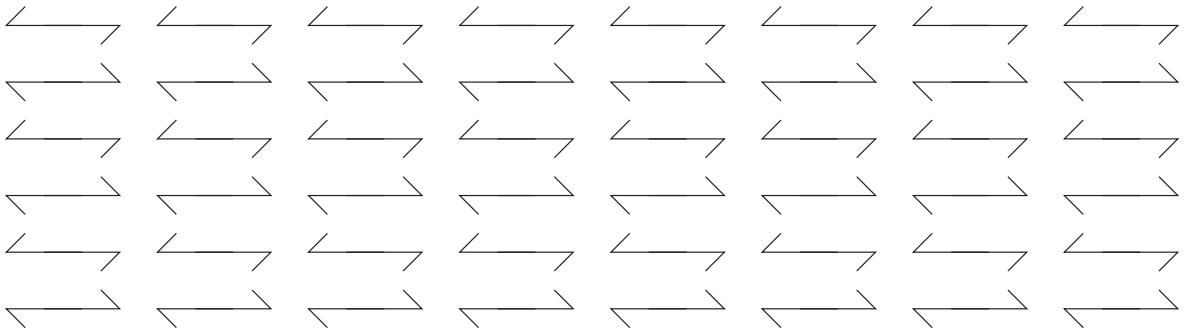
(a) First, assume that G contains a reflection σ_m such that the condition (2) of the previous lemma is satisfied. The lattice determined by the two generating translations from the lemma is rhombic. Let us prove that line m must contain a center of a halfturn. Consider a rhombus that is intersected by m, whose corners are centers of halfturns and whose interior does not contain any such centers. We know that m is parallel to a diagonal of the rhombus. If m is not the diagonal then one of the corners is mapped inside the rhombus by reflection σ_m, which contradicts the fact that there are no rotation centers inside the rhombus. We conclude that m bisects the rhombus along its diagonal, and therefore m contains a center of rotation. As the first halfturn σ_P was chosen arbitrarily, we can choose it in such a way that $P \in m$. We see that line m is then uniquely determined by τ_1, τ_2 and P. All other odd elements of G are then the products of σ_m and even isometries. This gives the wallpaper group $W_2^1 = \langle\tau_1, \tau_2, \sigma_P, \sigma_m\rangle$ where τ_1 and τ_2 are of equal length, and m is the line through P and $\tau_1\tau_2(P)$. This group is the symmetry group of



(b) Assume then that G contains a reflection σ_m that satisfies the condition (1) of the lemma. Let us call the direction of τ_1 and m the horizontal direction. We have two possibilities: (i) that m contains a center of a halfturn, and (ii) that m does not contain a center of a halfturn. In the second case the line m must run in the middle between two horizontal rows of rotation centers. As before, all other odd isometries are uniquely determined by σ_m and the even isometries. We get two wallpaper groups $W_2^2 = \langle \tau_1, \tau_2, \sigma_P, \sigma_m \rangle$ where m is the line through P and $\tau_1(P)$, and $W_2^3 = \langle \tau_1, \tau_2, \sigma_P, \sigma_m \rangle$ where m is the perpendicular bisector between points P and the center of halfturn $\tau_2\sigma_P$. In both cases, τ_1 and τ_2 are perpendicular. Group W_2^2 is the symmetry group of

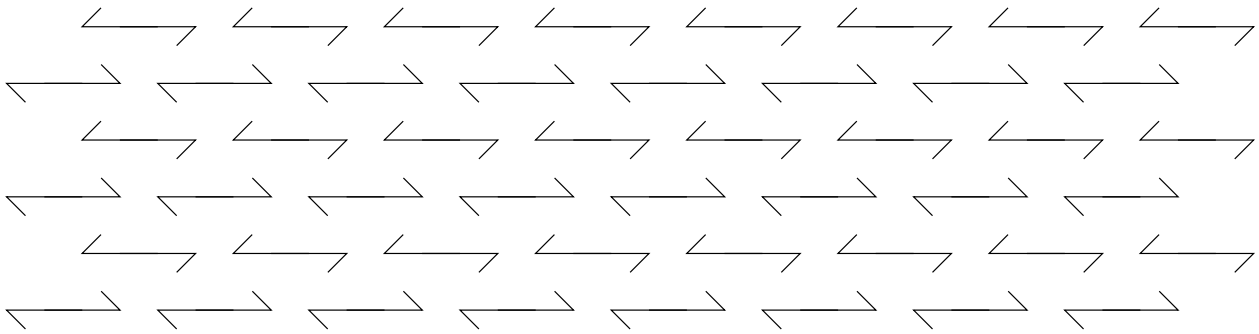


and group W_2^3 is the symmetry group of

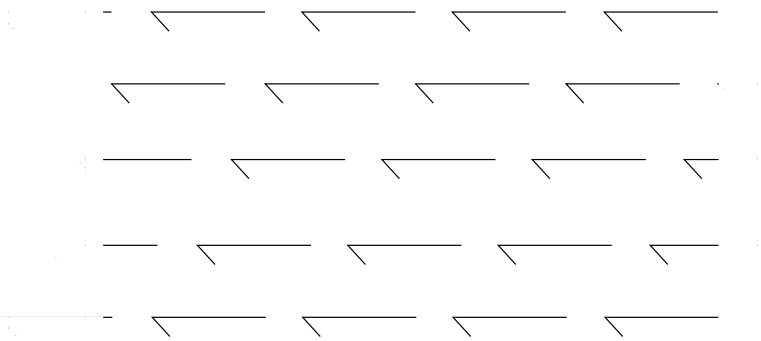


(c) Finally, assume that G does not contain any reflections. Let $\gamma \in G$ be a glide reflection with axis m. According to the last claim of Lemma 2.32, case (1) of the lemma must apply. Let τ_1 and τ_2 be two

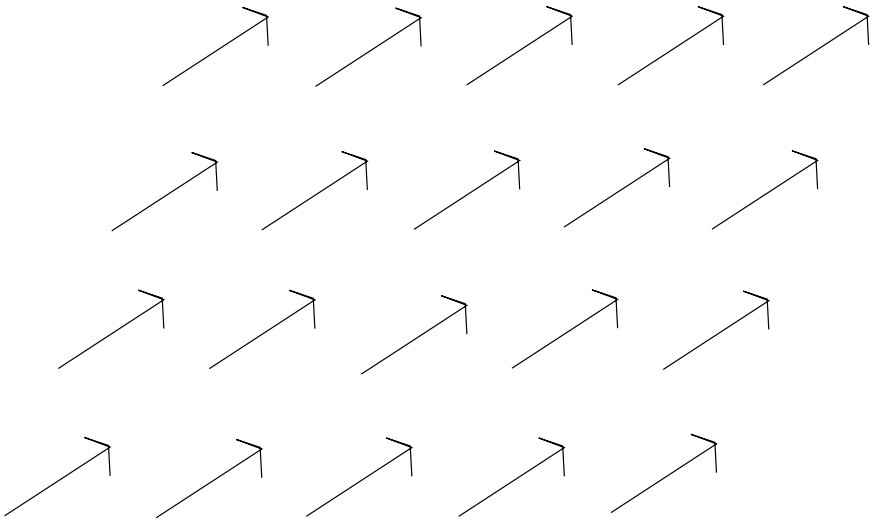
perpendicular translations, as indicated by the case (1) of the lemma. If the axis m contains the center P of some halfturn $\sigma_P \in G$ then G contains the reflection $\gamma\sigma_P$. We conclude that m must run in the middle between two horizontal rows of rotation centers. Let integer i be such that $\gamma^2 = \tau_1^i$. If i would be even then γ and τ_1 would generate a reflection, so i must be odd. By multiplying γ with a suitable power of τ_1 we obtain a glide reflection whose square is exactly τ_1. This is uniquely determined, so the group G is also determined. It is $W_2^4 = \langle \tau_2, \sigma_P, \gamma \rangle$ where γ is a glide reflection such that $\tau_1 = \gamma^2$ and τ_2 are perpendicular. This is the symmetry group of



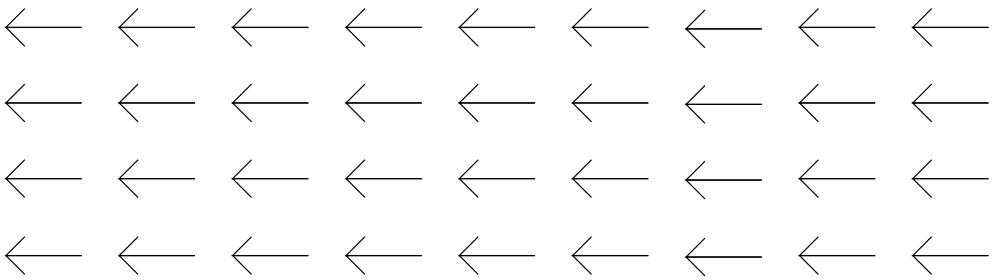
5) As our final case, assume that there are no non-trivial rotations in group G. The even isometries are then all translations generated by τ_1 and τ_2. If there are no odd isometries then the group is $W_1 = \langle \tau_1, \tau_2 \rangle$. This group is the symmetry group of



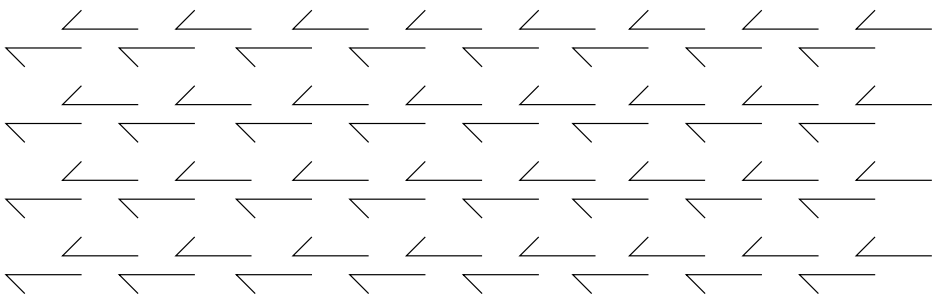
Let us assume then that G also contains odd isometries. If G contains a reflection σ_m then according to Lemma 2.32 either G has perpendicular generating translations τ_1 and τ_2 and m is parallel to τ_1, or G has generating translations τ_1 and τ_2 of equal length and m is parallel to $\tau_1\tau_2$. In the second case we obtain group $W_1^1 = \langle \tau_1, \tau_2, \sigma_m \rangle$ that is the symmetry group of



and in the first case we obtain the symmetry group $W_1^2 = \langle \tau_1, \tau_2, \sigma_m \rangle$ of



Assume then that G does not contain any reflections but contains a glide reflection with axis m. Case (1) of Lemma 2.32 must apply. Then we can choose the glide reflection γ in such a way that $\tau_1 = \gamma^2$. This gives the last wallpaper group $W_1^3 = \langle \gamma, \tau_2 \rangle$. A pattern with this symmetry group is for example



We have exhausted all possibilities of wallpaper groups. We found 17 groups: Two with 6-fold rotations, three with 4-fold rotations, three with 3-fold (but no 6-fold) rotations, five with halfturns (but no higher order rotations) and four without non-trivial rotations.

Theorem 2.33 *Let G be a wallpaper group. Then G is among the 17 groups discussed above.* □

2.11 Final remarks on discrete symmetry groups

The rosette groups, frieze groups and the wallpaper groups have standard names given by crystallographers, and standardized by the International Union of Crystallography. Another naming system was developed by Fejes Tóth. The following table summarizes these notations:

Our notation	Fejes Tóth	Crystallographic
C_n	C_n	n
D_n	D_n	nm, if m is odd, $n$mm, if m is even
F_{1001}	F_1^1	p1m1
F_{1111}	F_2^1	pmm2
F_{0000}	F_1	p111
F_{0100}	F_1^2	pm11
F_{0010}	F_2	p112
F_{0001}	F_1^3	p1a1
F_{0111}	F_2^2	pma2
W_6	W_6	p6
W_6^1	W_6^1	p6m
W_3	W_3	p3
W_3^1	W_3^1	p3m1
W_3^2	W_3^2	p31m
W_4	W_4	p4
W_4^1	W_4^1	p4m
W_4^2	W_4^2	p4g
W_2	W_2	p2
W_2^1	W_2^1	cmm
W_2^2	W_2^2	pmm
W_2^3	W_2^3	pmg
W_2^4	W_2^4	pgg
W_1	W_1	p1
W_1^1	W_1^1	cm
W_1^2	W_1^2	pm
W_1^3	W_1^3	pg

Observe that each rosette, frieze or wallpaper group type is actually a family of subgroups of $\mathcal{I}$. For example, for each $P \in \mathbb{R}^2$, the halfturn around point P generates the cyclic group C_2, but of course each choice of P provides a distinct subgroup of $\mathcal{I}$. In fact, each group type represents a family of affinely conjugate subgroups, as explained briefly below:

- An affine transformation of the plane is a transformation that preserves parallelism of lines. It is the composition of a linear transformation and a translation, that is, a mapping

$$f : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto M \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix}$$

  where M is a 2×2 matrix. The transformation is one-to-one if and only if M is invertible, i.e., $\det(M) \neq 0$. Isometries are exactly the distance preserving affine maps. Distance preservation is equivalent to M being an orthogonal matrix, i.e., equivalent to $MM^T = I$ where M^T is the transpose of M and I is the 2×2 identity matrix. Even and odd isometries correspond to orthogonal matrices M whose determinant is +1 and -1, respectively.

- Two subgroups G_1 and G_2 of $\mathcal{I}$ are said to be equal up to affine conjugacy if there exists a one-to-one affine transformation f such that $G_1 = fG_2f^{-1}$, that is, elements of G_1 are exactly the functions $f\alpha f^{-1}$ for $\alpha \in G_2$. (In particular, this requires that $f\alpha f^{-1}$ are isometries for all $\alpha \in G_2$, which is not the case for all affine f and all isometries $\alpha \in \mathcal{I}$.)
- If G_1 and G_2 are wallpaper groups, frieze groups or rosette groups then equality up to affine conjugacy exactly means that they are of the same wallpaper, frieze or rosette group type.
- Affine conjugacy preserves isometry types: If α and $f\alpha f^{-1}$ are both isometries then they are of the same type: both translations, both rotations, both reflections or both glide reflections. (To see this, note that the parity of the isometry is preserved by affine conjugacy, and that P is a fixed point of α if and only if $f(P)$ is a fixed point of $f\alpha f^{-1}$.) But as mentioned above, $f\alpha f^{-1}$ may also not be an isometry.
- As groups, C_2 and D_1 are isomorphic. But they are not equal up to affine conjugacy. Likewise, frieze groups F_{0000} and F_{0001} are isomorphic (both are infinite cyclic groups, one is generated by a translation the other one by a glide reflection) but we consider them different as they are not affinely conjugate.

3 Tilings

Intuitively, a tiling is a covering of the plane without overlaps using some tiles. We start by giving more precise definitions. You may want to review some basic concepts of topology (especially the standard Euclidean topology of $\mathbb{R}^2$) such as

- open and closed sets,
- neighborhood of a point (=any open set containing the point),
- interior of a set (=largest open set contained in the set),
- closure of a set (=smallest closed set containing the set),
- boundary of a set (=intersection of the closures of the set and its complement),
- compactness,
- continuity of functions (inverse images of open sets are open),
- homeomorphism (=continuous bijection whose inverse is also continuous).
- connectedness (a set is connected iff it is not the union of two disjoint open sets),

Recall that since the Euclidean topology of $\mathbb{R}^2$ is metric, it is Hausdorff, and compactness is equivalent to being closed and bounded. Also, in $\mathbb{R}^2$ an open set is connected if and only if it is path-connected, that is, each pair of its points can be joined by a path (=homeomorphic image of the unit interval) inside the set. Let us denote by

$$B_r(P) = \{X \in \mathbb{R}^2 \mid d(X, P) < r\}$$

the open disk of radius r centered at P, and if P is the origin O, we simply denote $B_r = B_r(O)$. The closure of an open disk is a closed disk

$$\overline{B}_r(P) = \{X \in \mathbb{R}^2 \mid d(X, P) \leq r\},$$

and $\overline{B}_r = \overline{B}_r(O)$.

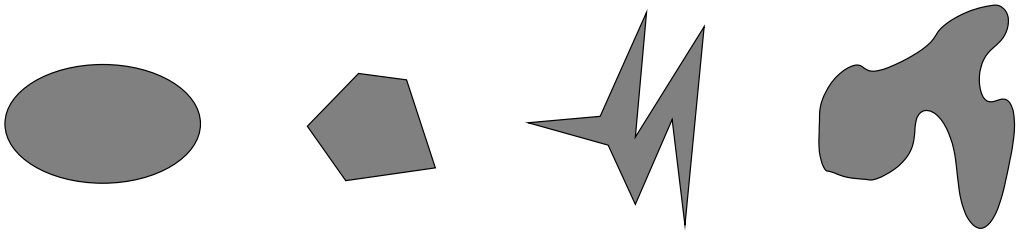
3.1 Basic definitions

A tile is a subset of $\mathbb{R}^2$ that is a topological disk. This means that it is the image of the closed disk $\overline{B}_1$ under some homeomorphism. Homeomorphisms preserve topological properties, so tile t immediately inherits topological properties from the disk $\overline{B}_1$:

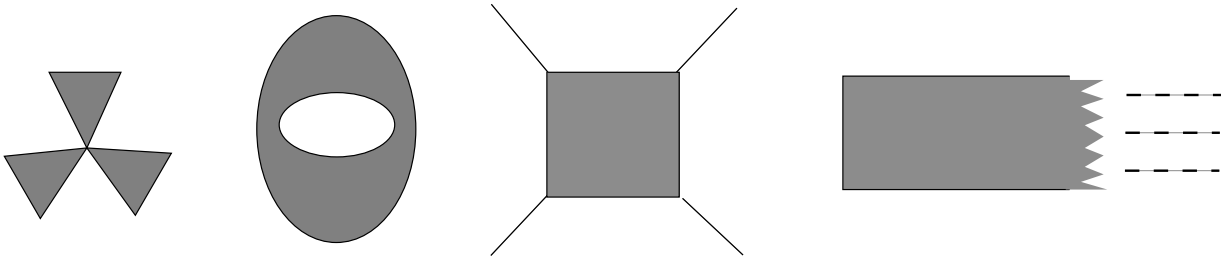
- t is compact (=closed and bounded),
- the interior of t is connected, and the complement of t is connected,
- the boundary of t is the boundary of its interior,
- the boundary of t is a simple closed curve, that is, homeomorphic to the unit circle

$$\{X \in \mathbb{R}^2 \mid d(X, O) = 1\}.$$

This definition of a tile is very general. Later, additional restrictions will be added as needed. For example, we may restrict our attention to tiles that are polygons. Here are some examples of tiles:



but these are not tiles:



(They are with non-connected interior, non-connected complement, boundary that is not the boundary of the interior, and unbounded, in this order.)

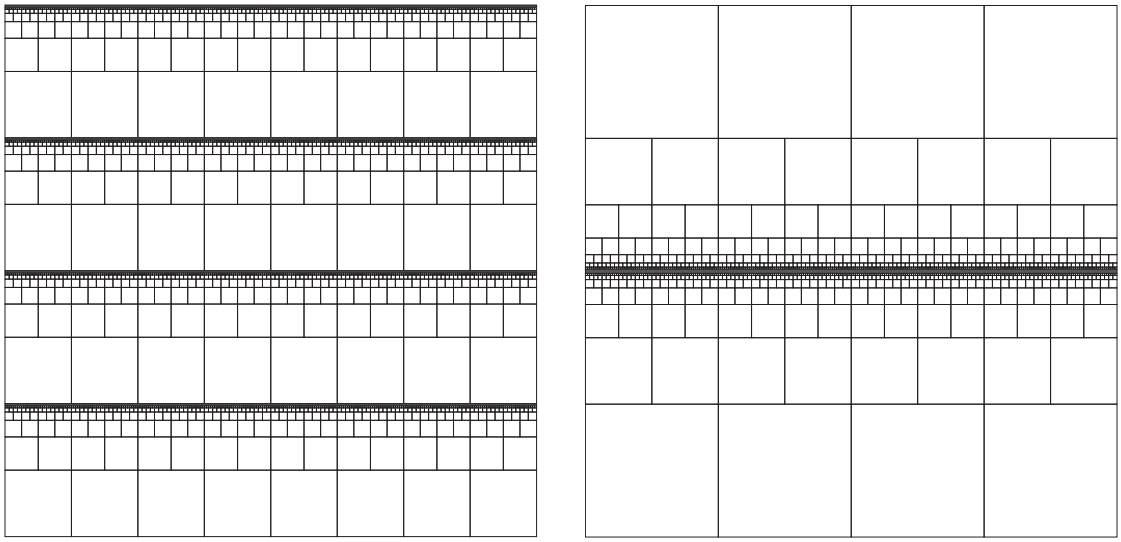
A tiling $\mathcal{T}$ is a family of tiles that covers the plane

(1) without gaps (every $P \in \mathbb{R}^2$ belongs at least one tile), and

(2) without overlaps (the interiors of the tiles are pairwise disjoint).

Notice that the boundaries of the tiles do not need to be disjoint. But it follows that every point that belongs to more than one tile cannot belong to the interior of any tile. Notice also that the number of tiles in any tiling must be infinite (union of a finite number of bounded sets would be bounded) but countable (the interior of each tile contains a point with rational coordinates).

This definition of tilings is very general. It does not restrict the number of different shapes used in any way, so one tiling can, for example, contain arbitrarily small tiles. The left picture below represents a part of a tiling, while the rightmost picture is not a tiling since the horizontal line in the center is not covered by any tile.



Let $\mathcal{T} = \{t_1, t_2, \dots\}$ be a tiling. Its symmetry group G consists of those isometries α that take every tile of $\mathcal{T}$ onto a tile of $\mathcal{T}$, that is, for every $i = 1, 2, \dots$ there exists j such that $\alpha(t_i) = t_j$. It is easy to see that symmetry groups of tilings (even under our very general definition of tiles) are discrete: the only possibilities are our familiar rosette, frieze and wallpaper groups.

Theorem 3.1 *The symmetry group of a tiling is discrete.*

Proof. Let G be the symmetry group of tiling $\mathcal{T} = \{t_1, t_2, \dots\}$. Then there must exist a positive number ε such that the length of every non-trivial translation in G is at least ε. Indeed, the interior of tile t_1 contains a disk $B_\varepsilon(P)$ for some $\varepsilon > 0$, so any translation τ that is shorter than ε takes P into the interior of t_1. This means that $\tau(t_1) = t_1$, which is possible only if $\tau = \iota$.

Consider then rotations. Suppose first there is a non-trivial translation τ in G. If there are arbitrarily small rotations in G then there are arbitrarily small translations among $\tau^{-1}\rho\tau\rho^{-1}$, which contradicts the conclusion in the previous paragraph.

Suppose than that G contains only the trivial translation. Then all rotations have the same center P of rotation (Corollary 2.22). Suppose there would be arbitrarily small rotations around P.

Let $t \in \mathcal{T}$ be a tile that contains point P. We have $t \subseteq B_k(P)$ for a sufficiently large number k. Let Q be a point whose distance from P is at least k such that Q belongs to the interior of some tile $t' \in \mathcal{T}$. (Just choose any point Q sufficiently far away from P. If Q is not in the interior of any tile then Q is on the boundary of some t'. There are interior points of t' close to Q. We can choose any one of them.)

The circle $c = \{X \in \mathbb{R}^2 \mid d(P, X) = d(P, Q)\}$ does not intersect t, but it contains an interior point Q of t'. Let us prove that $c \subseteq t'$. Assume the contrary: there exists a point $R \in c$ such that $R \notin t'$. The complement of t' is open so, for all sufficiently small angles Θ, we have $\rho_{P,\Theta}(R) \notin t'$.

Let $\varepsilon > 0$ be a small number so that $\rho_{P,\Theta}(Q)$ is an interior point of t' and $\rho_{P,\Theta}(R) \notin t'$ for all angles Θ with $|\Theta| < \varepsilon$. Choose one positive angle $\Theta < \varepsilon$ such that $\rho = \rho_{P,\Theta} \in G$. Because ρ is a symmetry of the tiling such that $\rho(Q)$ is an interior point of t', we must have that $\rho(t') = t'$. This means that $\rho^i(Q) \in t'$ for all integers i. Choose number i such that $|i\Theta - \Phi| < \varepsilon$ where Φ is the angle such that $\rho_{P,\Phi}(Q) = R$. Then $\rho^i(Q) \in t'$ but, on the other hand,

$$\rho^i(Q) = \rho_{P,i\Theta}(Q) = \rho_{P,i\Theta-\Phi}\rho_{P,\Phi}(Q) = \rho_{P,i\Theta-\Phi}(R) \notin t',$$

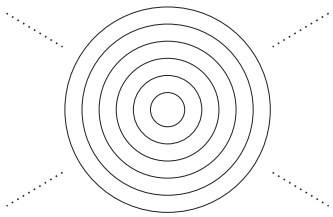
a contradiction.

We have proved that $c \subseteq t'$. Then the complement of t' is not connected: Interior points of t are in the disk $B_k(P)$ so they are separated by t' from the points outside the circle c. This contradicts the fact

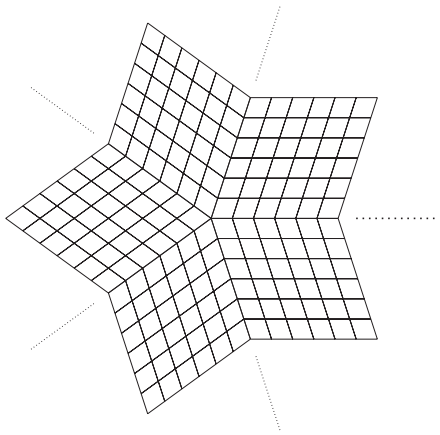
that t' should be a topological disk. Conclusion: there can only be a finite number of rotations in G, so G is a finite subgroup of $\mathcal{I}$, and therefore a rosette group.

□

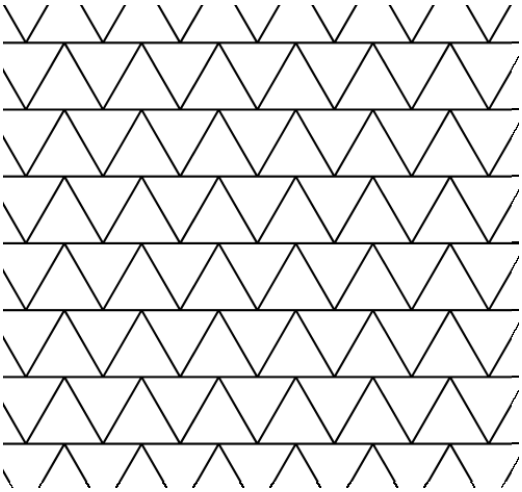
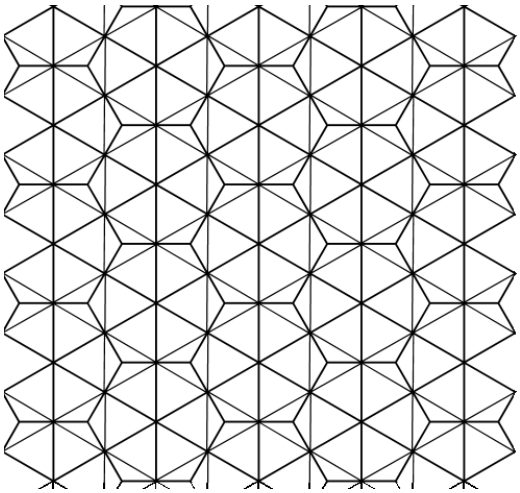
Note that it is essential in the proof that the tiles are topological disks, and hence do not contain holes. If we would allow tiles that are topological rings then we would have, for example, the following "tiling" whose symmetry group is not discrete.



Each rosette group, frieze group and wallpaper group is the symmetry group of some tiling. We see some examples in the homeworks. As another example, below is a piece of a tiling whose symmetry group is D_5. This can be easily generalized to obtain a tiling whose symmetry group is D_n or C_n, for any $n \geq 5$.



Our main interest is in tilings using only a finite number of different shapes. More precisely, tiles $\{p_1, p_2, \ldots, p_k\}$ are prototiles of a tiling $\mathcal{T} = \{t_1, t_2, \ldots\}$ if every tile $t_i \in \mathcal{T}$ is congruent to some p_j. By congruent we mean that there is an isometry (even or odd!) that takes t_i onto p_j. We say that the prototiles $\{p_1, p_2, \ldots, p_k\}$ admit the tiling $\mathcal{T}$. Tiling $\mathcal{T}$ is called k-hedral, where k is the number of prototiles p_j. In the special cases of $k = 1$ and $k = 2$ the tiling is called monohedral and dihedral, respectively. Note that some tiles may be "flipped over" copies of the prototiles, that is, the isometry that takes the prototile on a tile may be odd. In some cases we may be interested in those k-hedral tilings where the tiles are congruent to prototiles by even isometries, but in these cases this will be stated explicitly. Here is an example of a monohedral and a dihedral tiling:

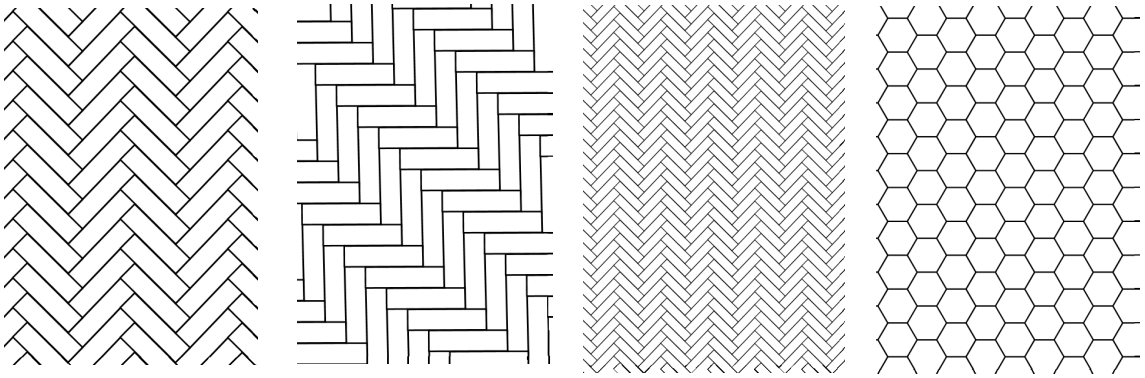



Let $\mathcal{T} = \{t_1, t_2, t_3, \dots\}$ be a tiling. If $h : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is a homeomorphism then also $h(\mathcal{T}) = \{h(t_1), h(t_2), h(t_3), \dots\}$ is a tiling. We say that tilings $\mathcal{T}$ and $h(\mathcal{T})$ are topologically equivalent. This is easily seen to be an equivalence relation among tilings.

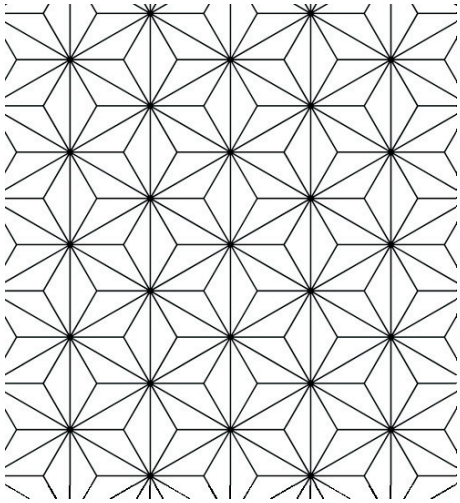
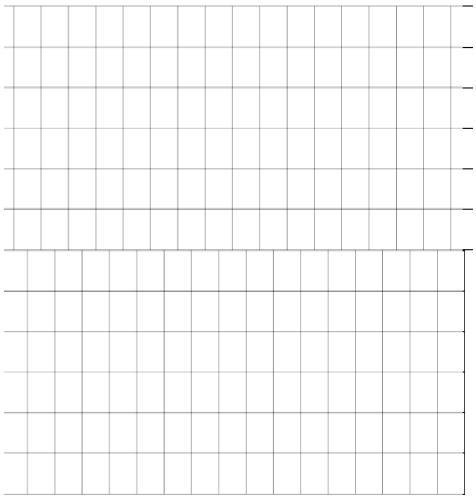
Every isometry is a homeomorphism, so if α is an isometry then $\alpha(\mathcal{T}) = \{\alpha(t_1), \alpha(t_2), \alpha(t_3), \dots\}$ is a tiling. We say that that $\alpha(\mathcal{T})$ is congruent to tiling $\mathcal{T}$. Also congruence is an equivalence relation among tilings.

Finally, a similarity $s : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is a composition of an isometry and a stretch (that is, a function that maps $(x, y) \mapsto (kx, ky)$ for some $k > 0$). In other words, a similarity s by factor $k > 0$ is a function such that for any two points $P, Q \in \mathbb{R}^2$ we have $d(s(P), s(Q)) = k \cdot d(P, Q)$. Similarities are homeomorphisms, so $s(\mathcal{T}) = \{s(t_1), s(t_2), s(t_3), \dots\}$ is a tiling. We say that tilings $\mathcal{T}$ and $s(\mathcal{T})$ are similar. Intuitively, similarity of two tiling means that they look the same when one of them is watched under a suitable magnifying class. Usually (unless otherwise noted) we consider similar tilings to be the same tiling.

The following figure contains four topologically equivalent monohedral tilings. First two are congruent with each other, and they are similar to the third one:



Two tiles t_1 and t_2 of tiling $\mathcal{T}$ are called equivalent in $\mathcal{T}$ if there exists a symmetry of $\mathcal{T}$ that takes t_1 onto t_2. This is clearly an equivalence relation among tiles t_i. Equivalence classes are called the transitivity classes of $\mathcal{T}$. If tiling $\mathcal{T}$ has only one transitivity class then the tiling is called isohedral (or tile-transitive). More generally, if there are k transitivity classes then the tiling is called k-isohedral. Notice that any isohedral tiling is monohedral as equivalent tiles are congruent. But there are monohedral tilings that are not isohedral. Analogously, a k-isohedral tiling is always k-hedral (but it can also be n-hedral for some $n < k$). Here are examples of an isohedral tiling and a monohedral tiling that is not isohedral, or even k-isohedral for any finite k.

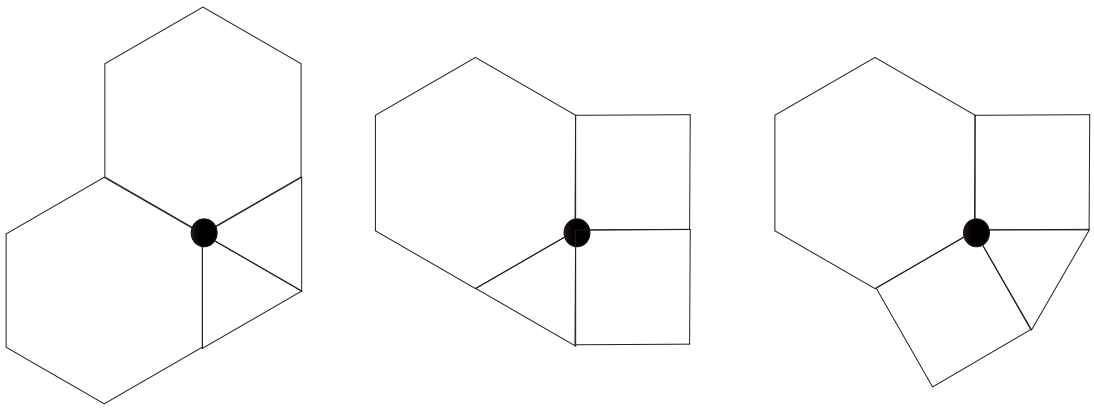



It is easy to see (in the homeworks!) that the symmetry group of a k-hedral tiling is a wallpaper group if and only if the tiling is n-isohedral for some n. (But there are also tilings that are not k-hedral for any k and whose symmetry group is a wallpaper group.)

3.2 Tilings by regular polygons

We restrict the study in this section to tilings that are by regular polygons, and that are edge-to-edge, that is, the intersection of two tiles is either empty, single vertex of the polygons, or the entire edge of the two neighboring polygons. Two tiles are called edge neighbors (vertex neighbors) if their intersection is an edge (edge or vertex, respectively) of the polygons. Corners of the polygons are called the vertices of the tiling.

Consider a vertex P where r regular polygons of orders $n_1, n_2, n_3, \ldots, n_r$ meet, in this order (counted clockwise or counterclockwise). Then we say that the vertex is of type $n_1 \cdot n_2 \cdot \cdots \cdot n_r$. For example, vertices of types $3 \cdot 3 \cdot 6 \cdot 6$, $3 \cdot 4 \cdot 4 \cdot 6$ and $3 \cdot 4 \cdot 6 \cdot 4$ look like



Notice that types $3{\cdot}4{\cdot}4{\cdot}6$ and $4{\cdot}6{\cdot}3{\cdot}4$ and $4{\cdot}3{\cdot}6{\cdot}4$ are all identical, as they are obtained by changing the starting point and/or the direction of reading the polygons. We also adapt the usual shorthand notations for repetitions, so that $3 \cdot 3 \cdot 6 \cdot 6$ may be abbreviated as $3^2 \cdot 6^2$.

The interior angle of a regular n-gon is $180^\circ(1-\frac{2}{n})$. Consequently, if P is a vertex of type $n_1 \cdot n_2 \cdot \cdots \cdot n_r$ then

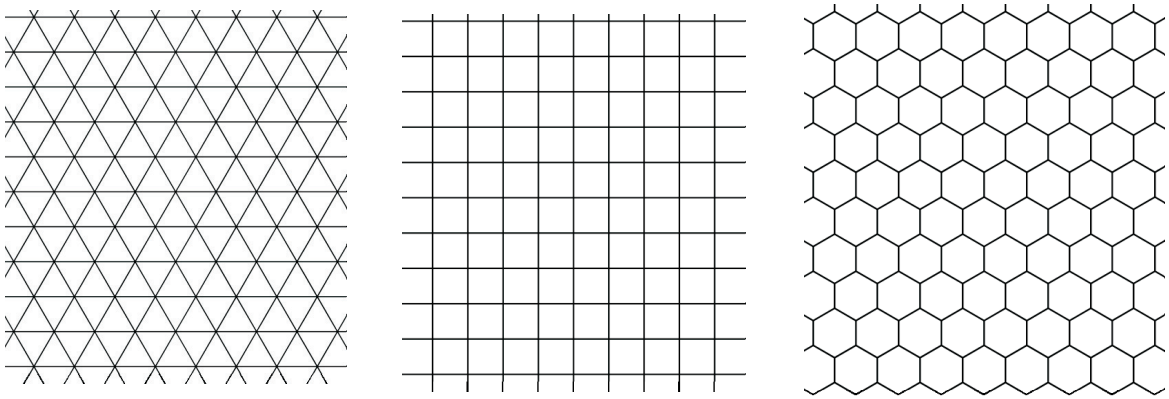
$$\sum_{i=1}^{r} \left(1 - \frac{2}{n_i}\right) = 2. \tag{1}$$

This follows from the fact that the interior angles of the polygons that meet at P must sum up to $360°$.

Assume first that the tiling is monohedral, with all tiles regular n-gons. Then (1) becomes

$$r(1 - \frac{2}{n}) = 2,$$

which implies $n = \frac{2r}{r-2}$. Because n is positive, we must have $r \geq 3$, and because $n \geq 3$ we must have $r \leq 6$. With $r = 3, 4, 5$ and 6 we get $n = 6, 4, \frac{10}{3}$ and 3. Number n is an integer so we only have three solutions. These are the familiar regular tilings



Theorem 3.2 *The only edge-to-edge monohedral tilings by regular polygons are the three regular tilings above.* □

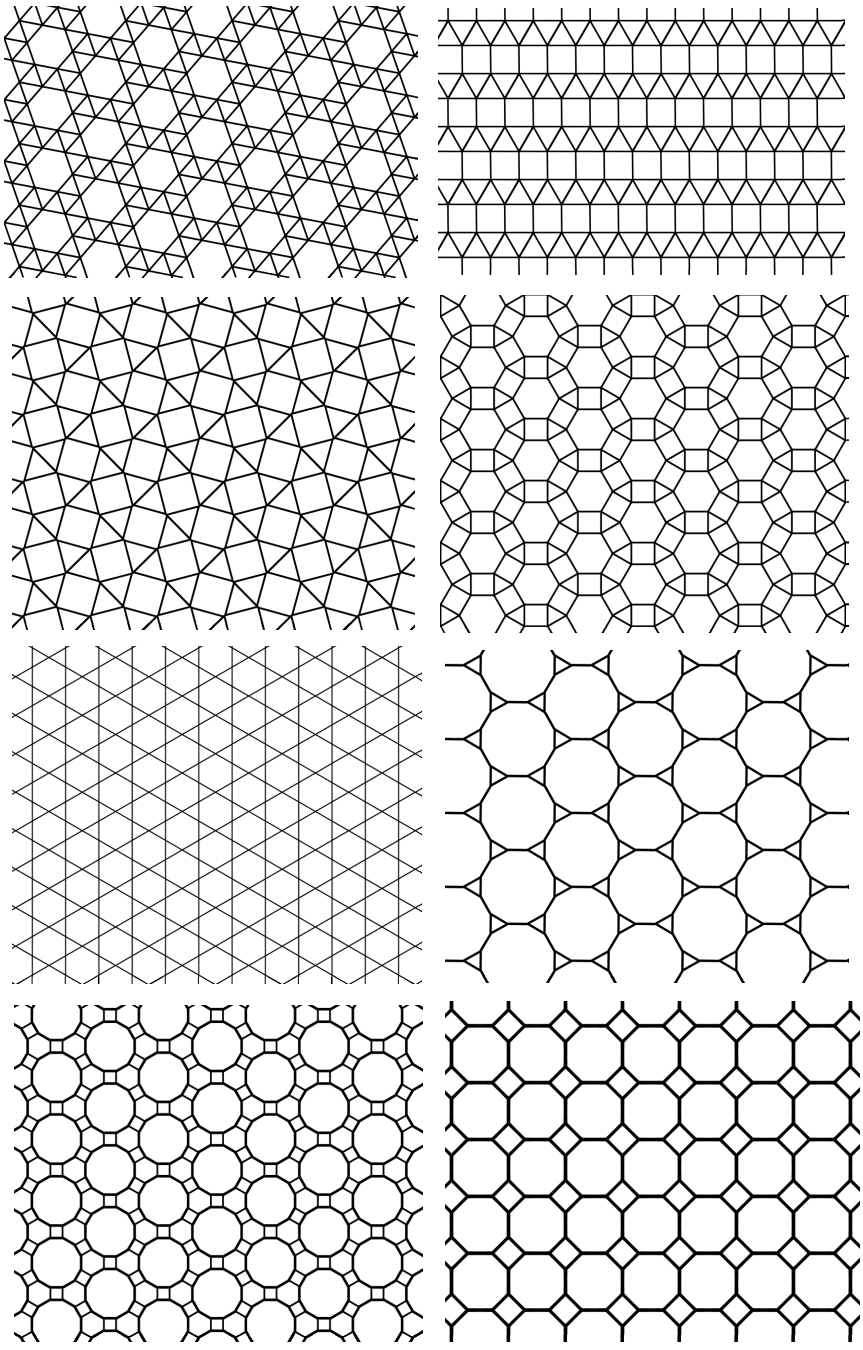
Consider then the case when the tiling is not necessarily monohedral. Possible types of vertices are limited by (1). We only have the following numerical solutions to (1), and the corresponding possibilities for the vertex types:

type	archimedean
$3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3$	A
$3 \cdot 3 \cdot 3 \cdot 3 \cdot 6$	A
$3 \cdot 3 \cdot 3 \cdot 4 \cdot 4$	A
$3 \cdot 3 \cdot 4 \cdot 3 \cdot 4$	A
$3 \cdot 3 \cdot 4 \cdot 12$	
$3 \cdot 3 \cdot 6 \cdot 6$	
$3 \cdot 4 \cdot 3 \cdot 12$	
$3 \cdot 4 \cdot 4 \cdot 6$	
$3 \cdot 4 \cdot 6 \cdot 4$	A
$3 \cdot 6 \cdot 3 \cdot 6$	A
$3 \cdot 7 \cdot 42$	
$3 \cdot 8 \cdot 24$	
$3 \cdot 9 \cdot 18$	
$3 \cdot 10 \cdot 15$	
$3 \cdot 12 \cdot 12$	A
$4 \cdot 4 \cdot 4 \cdot 4$	A
$4 \cdot 5 \cdot 20$	
$4 \cdot 6 \cdot 12$	A
$4 \cdot 8 \cdot 8$	A
$5 \cdot 5 \cdot 10$	
$6 \cdot 6 \cdot 6$	A

The last column indicates whether the vertex type appears in some archimedean tiling: An edge-to-edge tiling by regular polygons is termed archimedean if all vertices of the tiling are of the same type. The three regular tilings are all archimedean, corresponding to vertex types 6^3, 4^4 and 3^6. In addition, it turns out that there are only eight other examples of archimedean tilings, corresponding to the vertex types marked by "A" in the table above.

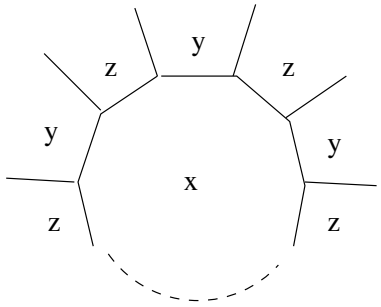
Theorem 3.3 (Kepler 1619) *There are exactly eleven different archimedean tilings, one of each type indicated by "A" in the table above.*

Proof. The eight non-regular archimedean tilings are shown below. It is easy to very that they are indeed archimedean, and one can easily verify that the types of their vertices match the types marked by "A".



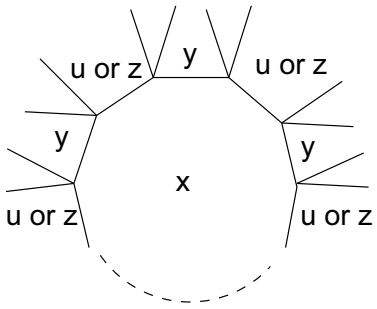
To prove that no other archimedean tilings exist we have to show that (i) the vertex types without "A" in the table are not possible in archimedean tilings, and (ii) each type with "A" leads to a unique tiling. Let us use the following terminology: a polygon is incident to its vertices and edges, and an edge is incident to its endpoints. Two vertices are adjacent if they are the two endpoints of an edge.

(i) Vertex type $x \cdot y \cdot z$ where x is odd and $y \neq z$ is not possible in any archimedean tiling: The edge neighbors of an x-gon across two consecutive edges are a y-gon and a z-gon. (Note: This is true even if $x = y$ or $x = z$.) So y-gons and z-gons alternate as the edge neighbors of an x-gon when we go around its edges clockwise. But since x is odd this is not possible: we necessarily end up with two consecutive neighbors of the same type.



This reasoning rules out six vertex types $3 \cdot 7 \cdot 42$, $3 \cdot 8 \cdot 24$, $3 \cdot 9 \cdot 18$, $3 \cdot 10 \cdot 15$, $4 \cdot 5 \cdot 20$ and $5 \cdot 5 \cdot 10$.

By a similar argument, vertex type $x \cdot y \cdot u \cdot z$ is not possible when x is odd, $y \neq z$, and no three of the numbers are equal. Clearly $x \neq y$ or $x \neq z$. The two situations are symmetric, so we may assume that $x \neq z$. Then two consecutive edge neighbors of an x-gon are an y-gon and a z-gon, or — if $x = y$ — possibly a y-gon and a u-gon. In either case, every other edge neighbor is a y-gon, and every other neighbor is not a y-gon, which is not possible as x is odd.

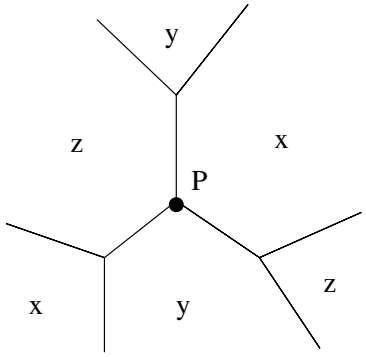


This rules out the remaining four vertex types $3 \cdot 3 \cdot 4 \cdot 12$, $3 \cdot 3 \cdot 6 \cdot 6$, $3 \cdot 4 \cdot 3 \cdot 12$ and $3 \cdot 4 \cdot 4 \cdot 6$.

(ii) Let us prove that any archimedean tiling $\mathcal{T}$ is similar to one of the given eleven tilings, namely the one with the same vertex type. We start by selecting one arbitrary vertex P of $\mathcal{T}$ and one arbitrary vertex P' of the known archimedean tiling $\mathcal{A}$ of the correct vertex type. There clearly exists a similarity function s that maps P onto P' in such a way that the polygons incident to P in $\mathcal{T}$ are mapped onto the polygons incident to P' in $\mathcal{A}$. Let us show that (with one exception in type $3 \cdot 3 \cdot 3 \cdot 3 \cdot 6$) similarity s maps the entire tiling $\mathcal{T}$ onto tiling $\mathcal{A}$.

It is enough to consider the vertices that are adjacent to P, and to show that all tiles incident to those vertices are mapped by s onto similar tiles on tiling $\mathcal{A}$. Namely then we can repeat the reasoning on the adjacent vertices to conclude that all vertices adjacent to them are mapped correctly, and so on, by mathematical induction, that all tiles at any distance from P are mapped onto tiles of $\mathcal{A}$.

Consider first tilings of vertex types $3\cdot12\cdot12$, $4\cdot6\cdot12$, $4\cdot8\cdot8$, and $6\cdot6\cdot6$, that is, the cases $x\cdot y\cdot z$ where three polygons meet at the vertices. Let Q be any of the three vertices adjacent to P. Two polygons incident to Q are also incident to P so they are known. This means that also the third polygon incident to Q is known and it must be mapped by s onto the corresponding tile in the archimedean tiling $\mathcal{A}$.

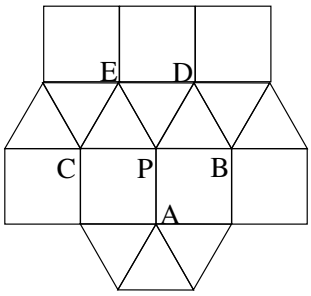


As discussed above, this is enough to prove that the entire tiling $\mathcal{T}$ is mapped onto $\mathcal{A}$.

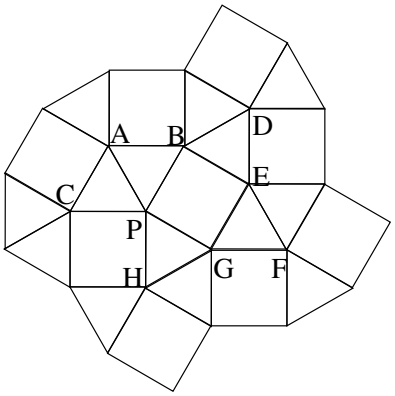
Vertex types $4\cdot4\cdot4\cdot4$ and $3\cdot3\cdot3\cdot3\cdot3\cdot3$ are also trivial: the polygons are all congruent and they must be correctly mapped onto the corresponding tiles in $\mathcal{A}$.

Consider then the vertex types $3\cdot4\cdot6\cdot4$ and $3\cdot6\cdot3\cdot6$. Let Q be a vertex adjacent to P. Two polygons that are edge neighbors and incident to Q are known. The other two are then also uniquely determined: in the first case one of the known polygons is a square, and the polygon opposite to it at Q must be a square as well, and in the case of $3\cdot6\cdot3\cdot6$ one of the known polygons is a triangle, and the polygon opposite to it at Q is a triangle. In both cases the polygons incident to Q are uniquely determined, and therefore mapped by s onto similar tiles in the tiling $\mathcal{A}$.

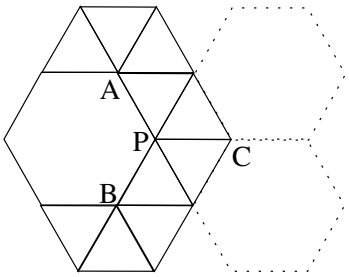
There remain three vertex types to analyze, namely $3\cdot3\cdot3\cdot3\cdot6$, $3\cdot3\cdot3\cdot4\cdot4$ and $3\cdot3\cdot4\cdot3\cdot4$. Consider type $3\cdot3\cdot3\cdot4\cdot4$ first: The following figure shows the order in which the vertices adjacent to P can be processed to determine the polygons incident to them. One can easily verify that the polygons are uniquely determined if the vertices are processed in the alphabetical order $A, B, C, D, \ldots$. So the tiles are all mapped correctly onto tiling $\mathcal{A}$.



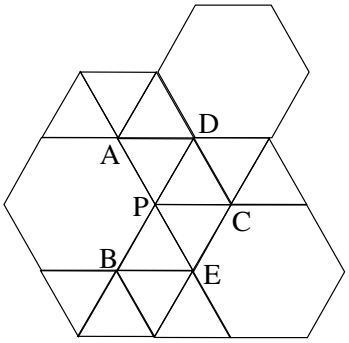
Analogously, if the vertex type is $3\cdot3\cdot4\cdot3\cdot4$ the vertices should be processed in the order indicated in this figure:



Finally, consider the vertex type $3 \cdot 3 \cdot 3 \cdot 3 \cdot 6$. In all previous cases, any similarity s that takes a vertex P and the incident polygons of $\mathcal{T}$ onto a vertex P' and its incident polygons of $\mathcal{A}$ is necessarily a similarity between entire tilings $\mathcal{T}$ and $\mathcal{A}$. But in the case of vertices of the type $3 \cdot 3 \cdot 3 \cdot 3 \cdot 6$ this is no longer true. Instead, there exist two similarities from vertex P onto vertex P': one even and one odd similarity. And exactly one of them is a similarity between tilings $\mathcal{T}$ and $\mathcal{A}$. In the following figure, the polygons incident to vertices A and B are uniquely determined. Then, the hexagon incident to vertex C must be one of the two dotted hexagons in the illustration. (The third alternative would lead to two hexagons that are vertex neighbors, and is therefore impossible.)



In either case, the similarity s can be chosen in such a way that the hexagon incident to C is mapped correctly onto tiling $\mathcal{A}$. The similarity is even or odd depending on the position of the hexagon. Thereafter, the remaining polygons are uniquely determined. In this case we have to verify the uniqueness of the polygons up to vertices of distance two from P. After this the uniqueness of the entire tiling follows by mathematical induction:



□

Notice that the previous proof indicates that the 11 archimedean tilings are vertex transitive: for any two vertices P_1 and P_2 of the tiling, there exists a symmetry of the tiling that takes P_1 onto P_2. With

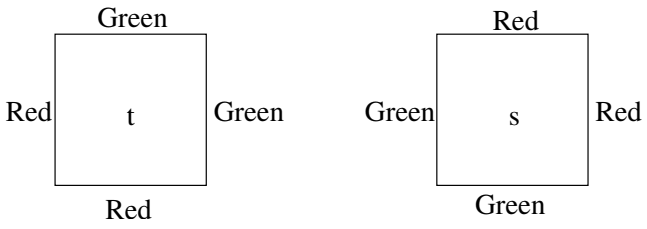
the exception of type $3 \cdot 3 \cdot 3 \cdot 3 \cdot 6$, any isometry that takes vertex P_1 and its incident polygons onto P_2 and its incident polygons is a symmetry of the tiling. The archimedean tiling of type $3 \cdot 3 \cdot 3 \cdot 3 \cdot 6$ comes in two enantiomorphic forms that are congruent with each other only by odd isometries.

Archimedean tilings are also called uniform, which refers to the fact that they are vertex transitive: the entire tiling looks exactly the same from each vertex. This is a stronger property than the property we started with: that the tiling looks locally the same at each vertex, as each vertex is of the same type.

4 Wang tiles

Wang tiles are unit square tiles with colored edges. Hence each tile can be represented as a 4-tuple (N, E, S, W) where N, E, S and W are the colors of the north, east, south and west sides of the square. Tilings with a finite number of prototiles are only considered. In Wang tilings copies of the prototiles are placed at integer lattice points, without rotating or flipping the tiles, so that all tiles are congruent to the given prototiles by translations only. A tiling can then be represented as a function $f : \mathbb{Z}^2 \to \mathcal{P}$ where $\mathcal{P}$ is the set of prototiles and $f(i, j)$ gives the tile at position $(i, j) \in \mathbb{Z}^2$. The tiling rule is that in a valid tiling the shared edge between any two tiles that are edge neighbors must have the same color.

For example, set $\mathcal{P} = \{$(Green, Green, Red, Red), (Red, Red, Green, Green)$\}$ consists of two prototiles



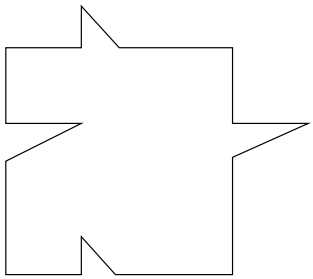
that admit the checkerboard-tiling

t	s	t	s	t	s	t
s	t	s	t	s	t	s
t	s	t	s	t	s	t
s	t	s	t	s	t	s

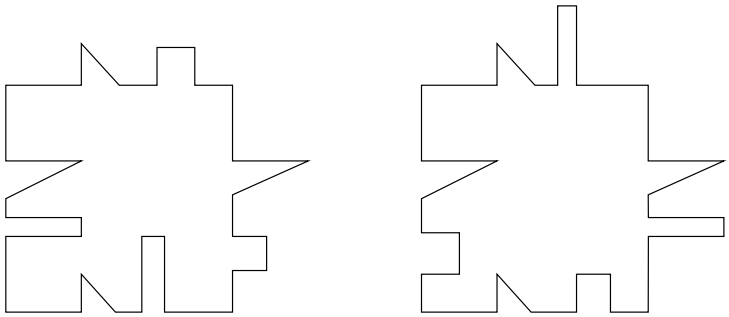
Wang tiles provide a discrete abstraction of tilings that allows us to study tilings using tools of discrete mathematics rather than geometry. This is especially useful when investigating computational properties and problems related to tilings. At first, Wang tiles may seem very restricted as the tilings are on a square lattice only. Nevertheless, the computational problems on Wang tiles are as hard as on more general types of tiles. By using Wang tiles we avoid problems related to representations of tiles (e.g. irrational coordinates of vertices) on computers, and we transform geometric problems into more manageable symbolic problems.

Our first observation is that Wang tiles fit our original definition of tiles as topological disks. We can namely represent Wang tiles as polygons as follows: The basic shape is a unit square. The middle of the north and east sides of each tile contain triangular "bumps" and the south and west sides have

"dents" that exactly fit the bumps. The bump/dent pairs are different in the horizontal and the vertical directions, and they are asymmetric so that flipped and non-flipped tiles do not match:



It should be clear that these tiles can only tile the plane in such a way that all tiles are aligned, and rotations and flips are not possible. To simulate the colors, we introduce an additional bump/dent pair on the sides of the tiles. Each color has its own bump/dent shape that does not fit with any other color. For example, our sample protoset of two tiles could look like this:



It should be obvious from this construction that any tiling by such polygons is congruent to a tiling where the tiles are positions at integer lattice points, without rotations and flips. Such tilings are clearly "isomorphic" to Wang tilings.

In the following we consider all tilings that given prototiles admit. In particular, we are interested to know when do given prototiles admit at least some tiling and when do they admit a periodic tiling. As it turns out that even among Wang tiles these questions can not be algorithmically answered (they are undecidable), so it follows that the questions are undecidable also among tiles that are polygons.

In the following two subsections we prove two preliminary results that will be needed in the algorithmic considerations that follow: First we show that if a finite set of Wang prototiles admits a tiling whose symmetry is a frieze group then it automatically admits also a tiling with a wallpaper symmetry. Then we prove that if one can tile arbitrarily large squares then one can also tile the entire infinite plane.

4.1 Periodic tilings

A tiling is called non-periodic if its symmetry group is finite, that is, if there is no translation that keeps the tiling invariant. A tiling is two-way periodic, or simply periodic, if its symmetry group is a wallpaper group, that is, if there are translations in non-parallel directions that keep the tiling invariant. A tiling whose symmetry group contains some non-trivial translation will be called called one-way periodic.

Vector $(a, b) \neq (0, 0)$ is called a period of a tiling, if $\tau_{(a,b)}$ is a symmetry of the tiling. In the case of a Wang tiling $f : \mathbb{Z}^2 \longrightarrow \mathcal{P}$ this means that $f(x, y) = f(x + a, y + b)$ for all $(x, y) \in \mathbb{Z}^2$.

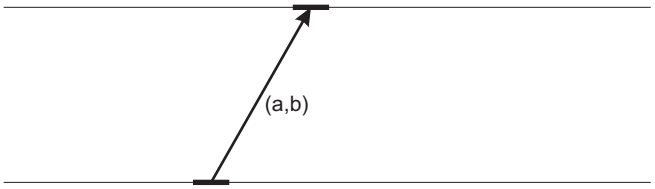
Notice that any two-way periodic tiling with Wang tiles has horizontal and vertical periods of equal lengths. Namely, if $f : \mathbb{Z}^2 \longrightarrow \mathcal{P}$ is periodic with non-parallel periods (a, b) and (c, d) then it is also

periodic with the horizontal period $d(a,b) - b(c,d) = (ad - bc, 0)$ and the vertical period $a(c,d) - c(a,b) = (0, ad - bc)$. Note that $ad - bc \neq 0$ as vectors (a,b) and (c,d) are not parallel. In other words, a two-way periodic Wang tiling consists of a periodic repetition of a square pattern.

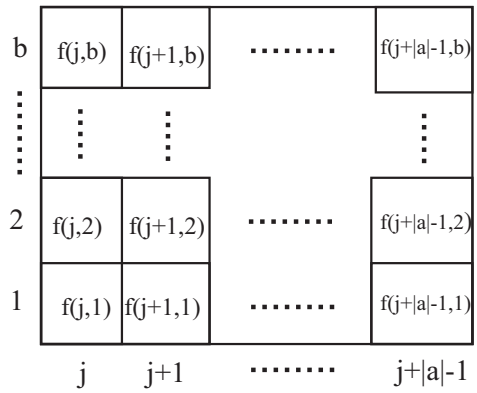
The next theorem states that a set of Wang tiles that admits a one-way periodic tiling also admits a periodic tiling:

Theorem 4.1 *Let $\mathcal{P}$ be a finite set of Wang prototiles that admits a tiling $f : \mathbb{Z}^2 \longrightarrow \mathcal{P}$ that is one-way periodic. Then there exists also a two-way periodic tiling $g : \mathbb{Z}^2 \longrightarrow \mathcal{P}$.*

Proof. Let $(a,b) \neq (0,0)$ be a period of tiling f. Without loss of generality we may assume that $b > 0$. Consider a horizontal strip of height b extracted from tiling f, e.g., the tiles $f(x,y)$ for $1 \leq y \leq b$. The sequences of horizontal colors on the top and the bottom of this strip are identical, with the horizontal offset a:

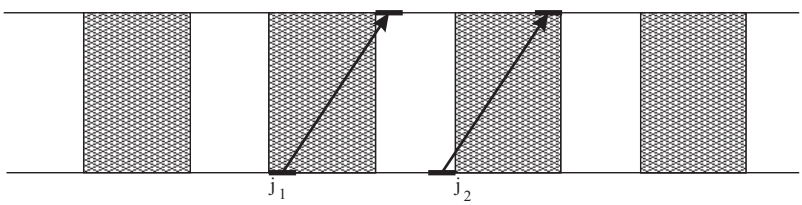


Within this strip, consider the rectangular $|a| \times b$ blocks

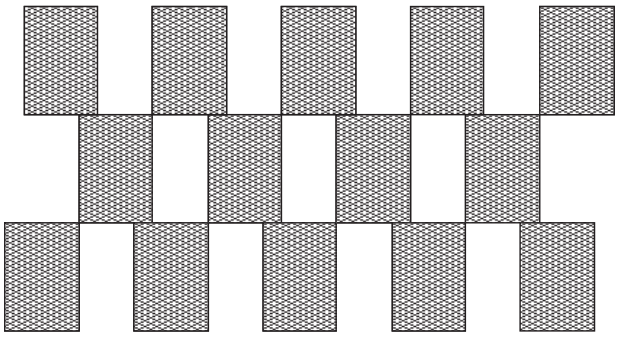


of tiles extracted from f with the bottom-left corner in position $(j,1)$, for all $j \in \mathbb{Z}$. (And if $a = 0$ consider just the sequences of vertical colors on the b rows.) Since there are only a finite number of tiles in the protoset, there are only a finite number of such blocks. This means that for two different values of j, say j_1 and j_2, the blocks are identical.

Now we can construct a valid periodic tiling of an infinite horizontal strip of height b by repeating the pattern between positions j_1 and j_2. Note that the sequences of horizontal colors on the top and the bottom of this strip are again identical, with the horizontal offset a:



The tiling of the strip is valid and it has a horizontal period of length $j_2 - j_1$. A valid, two-way periodic tiling of the plane can now be obtained by stacking copies of the strip on top of each other, with the horizontal offset a:



□

4.2 Compactness principle

In later chapters we'll introduce a metric on tilings that induces a compact topology. This will imply several interesting results, but for the main algorithmic questions that follow next we just need to know that if a Wang set admits tilings of arbitrarily large squares then it admits a tiling of the whole infinite plane. This is a direct consequence of the compactness, but we state the result here without a direct reference to topology.

Let $\mathcal{P}$ be a finite set of Wang prototiles. Let us call any function

$$c : \mathbb{Z}^2 \to \mathcal{P}$$

a *configuration*, and let us denote by

$$\mathcal{P}^{\mathbb{Z}^2} = \{c : \mathbb{Z}^2 \to \mathcal{P}\}$$

the set of all configurations over the tile set $\mathcal{P}$. Note that configurations are arbitrary assignments of tiles on integer lattice points. i.e., the color constraints are not checked. Valid tilings are particular types of configurations.

Consider an infinite sequence $c_1, c_2, \ldots$ of configurations, each $c_i \in \mathcal{P}^{\mathbb{Z}^2}$. We say that the sequence *converges* and $c \in \mathcal{P}^{\mathbb{Z}^2}$ is its *limit* if for every $(x, y) \in \mathbb{Z}^2$ there exists some $k \geq 0$ such that $c_i(x, y) = c(x, y)$ for all $i \geq k$. In other words: if we look at an arbitrary position and browse through a converging sequence $c_1, c_2, \ldots$ then from some moment on we always see the same tile in that position. It is obvious that if a limit exists it is unique, and we denote this limit by

$$\lim_{i \to \infty} c_i.$$

A *subsequence* of $c_1, c_2, \ldots$ is another sequence $c_{i_1}, c_{i_2}, \ldots$ where $i_1 < i_2 < \ldots$. A subsequence is hence obtained by picking infinitely many elements of the sequence, preserving their relative order. Obviously every subsequence of a converging sequence also converges and has the same limit.

The following theorem states the compactness of the configuration space:

Theorem 4.2 *Every sequence of configurations has a converging subsequence.*

Proof. Let $c_1, c_2, \ldots$ be an arbitrary sequence, $c_i \in \mathcal{P}^{\mathbb{Z}^2}$. Let $\vec{r}_1, \vec{r}_2, \ldots$ be some (arbitrary) enumeration of elements of $\mathbb{Z}^2$. In the following we show that there is a subsequence $c_{i_1}, c_{i_2}, \ldots$ such that for every

$n \geq 1$, if $j \geq n$ then $c_{i_j}(\vec{r}_n) = c_{i_n}(\vec{r}_n)$, i.e., the subsequence has a constant value in the n'th position $\vec{r}_n$ starting from the n'th element of the subsequence. Then clearly the subsequence converges.

Let us choose indices $i_0 < i_1 < i_2 < i_3 < \ldots$ inductively as follows: $i_0 = 0$ and $i_1 \geq 1$ is the smallest positive index such that there are infinitely many elements in $c_1, c_2, \ldots$ that agree with c_{i_1} in the first position $\vec{r}_1$. Such c_{i_1} exists because there are only finitely many different tiles that can appear in position $\vec{r}_1$.

Suppose then that i_{k-1} has been chosen and we want to choose i_k for $k \geq 2$. We choose i_k to be the smallest integer that satisfies the following three conditions:

(A_k) $i_k > i_{k-1}$,

(B_k) $c_{i_k}(\vec{r}_j) = c_{i_{k-1}}(\vec{r}_j)$ for all $j = 1, 2, \ldots k-1$.

(C_k) There exist infinitely many indices i such that $c_i(\vec{r}_j) = c_{i_k}(\vec{r}_j)$ for all $j = 1, 2, \ldots k$.

Numbers i_k that satisfy (A_k)–(C_k) always exist for the following reasons: Because condition (C_{k-1}) was satisfied when i_{k-1} was chosen, we have infinitely many choices of i_k that satisfy (B_k). Set $\mathcal{P}^k$ is finite so there is a finite number of combinations of tiles that can appear in positions $\vec{r}_1, \ldots, \vec{r}_k$. Consequently, among the infinitely many indices i_k that satisfy (B_k) there are infinitely many choices that also satisfy (C_k). Some of them hence satisfy all requirements (A_k)–(C_k).

It follows from properties (B_k) that $c_{i_1}, c_{i_2}, \ldots$ converges: For an arbitrary $\vec{r}_n \in \mathbb{Z}^2$ all c_{i_j} for $j \geq n$ have the same tile in position $\vec{r}_n$. □

Note: The proof is essentially the same as the proof of weak Kőnig's lemma which states that an infinite binary tree contains an infinite path. The proof did not require the axiom of choice. (The same result could also be easily proved using Tychonoff's theorem, but that is equivalent to the axiom of choice.)

Let us say that a configuration $c : \mathbb{Z}^2 \to \mathcal{P}$ tiles correctly at position $(x, y) \in \mathbb{Z}^2$ if $c(x, y)$ matches in color with its neighbors $c(x, y-1), c(x, y+1), c(x-1, y), c(x+1, y)$. A configuration is then a valid tiling iff it tiles correctly at each position.

The following corollary of the compactness principle states that if $\mathcal{P}$ can be used to properly tile arbitrarily large squares then it admits a valid tiling of the plane:

Corollary 4.3 *Let $\mathcal{P}$ be a finite set of Wang tiles. Suppose that for each finite set $F \subset \mathbb{Z}^2$ of positions there is a configuration that tiles correctly at each $(x, y) \in F$. Then $\mathcal{P}$ admits a valid tiling.*

Proof. Let $\vec{r}_1, \vec{r}_2, \ldots$ be an enumeration of elements of $\mathbb{Z}^2$, and for each $n \geq 1$ denote

$$F_n = \{\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_n\}.$$

By the hypotheses of the corollary there exists for each n a configuration c_n that tiles correctly at positions $\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_n$. By Theorem 4.2 the sequence $c_1, c_2, \ldots$ has a converging subsequence. Let $c \in \mathcal{P}^{\mathbb{Z}^2}$ be its limit. Then c tiles correctly at every position $\vec{r}_k$ because there are arbitrarily large indices i such that c and c_i assign the same tile to position $\vec{r}_k$ and its neighbors. □

4.3 Robinson's aperiodic tile set

It is easy to construct Wang tiles that admit non-periodic tilings. For a long time it was thought that any finite set of prototiles that admits a non-periodic tiling must also admit a periodic one. This conjecture was refuted by R.Berger in 1966 when he constructed a set of Wang prototiles that only admit non-periodic tilings.

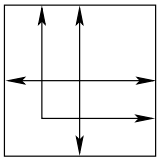
A finite set of prototiles is called aperiodic if

(i) it admits valid tilings, and

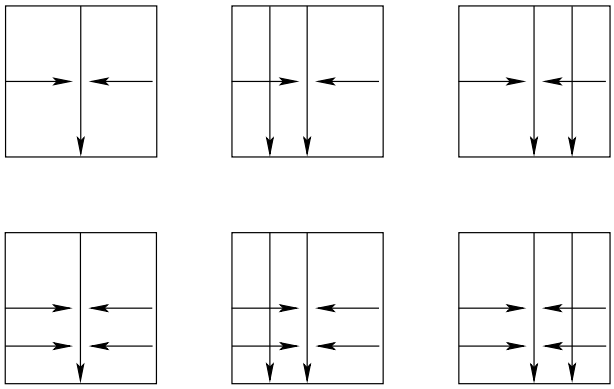
(ii) it does not admit any periodic valid tilings.

As an example of an aperiodic tile set we next describe a set of 56 Wang tiles due to R.M.Robinson. This set will be also useful later in our undecidability proofs. Instead of colors we use arrows to describe the matching rules between tiles. In valid tilings arrow heads and tails in neighboring tiles must match. This formalism can be easily converted into a color-based matching simply by assigning a different color for each orientation and positioning of arrows.

Robinson's tile set consists of tiles



called "crosses" and tiles



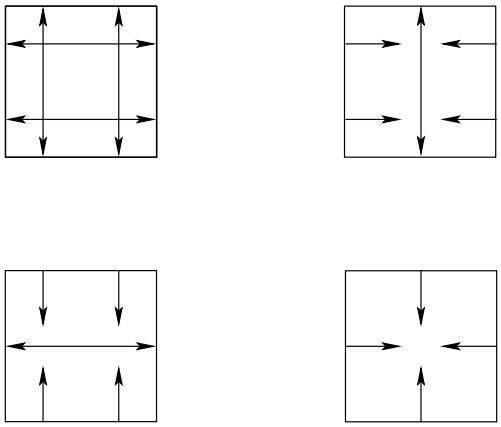
called arms. All tiles may be rotated so each tile comes in four orientations. Hence the total number of such tiles is 28.

The following terminology will be used:

- Every tile has central arrows at the centers of all four sides, and possibly some side arrows.
- A cross is said to face the directions of its side arrows.
- The arrow that runs through an arm is called the principal arrow of the arm, and the direction of the principal arrow is called the direction of the arm.

All six arms above are drawn in the north-to-south orientation. An important fact about arms is that if there are side arrows perpendicular to the principal arrow then these side arrows are towards the head of the principal arrow. Otherwise, all combinations of side arrows are allowed, as shown in the figure above.

We want to enforce a cross in the intersections of every other row and column. This can be established by forming the cartesian product ("sandwich tiles") with the parity tiles



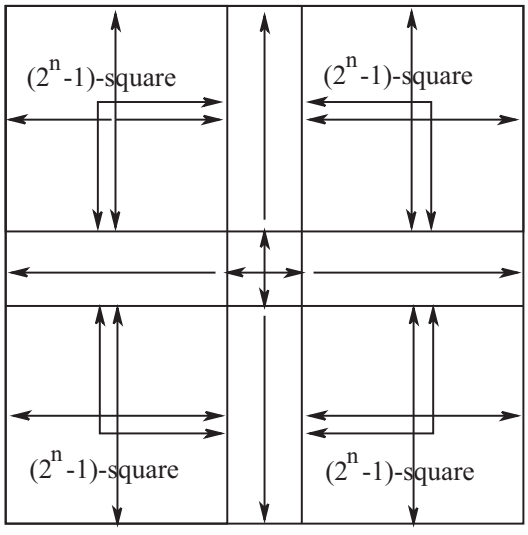
and by forbidding arms from the first parity tile. Since the only way the parity tiles tile the plane is by alternating the tiles on even and odd rows and columns, the first parity tile is forced at the intersections of every other row and column, and hence a cross is forced to appear in those locations. By numbering the rows and columns suitably we can assume from now on that all odd-odd positions of the plane contain a cross.

Note that between two crosses can only appear an arm, and the orientation of the arm has just two possible choices as it cannot point towards either cross. This means that the second parity tile only needs to be paired with north-to-south or south-to-north oriented arms, and the third parity tile is only paired with east-to-west or west-to-east oriented arms. The fourth parity tile is paired with any of the 28 tiles. So the final set contains $4 + 12 + 12 + 28 = 56$ different tiles.

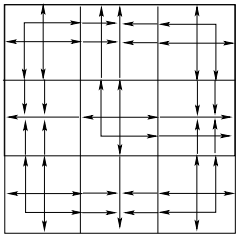
Next we investigate valid tilings admitted by Robinson's tiles, and we show that the tile set is aperiodic. Specific patterns called 1-, 3-, 7-, 15-, ..., $(2^n - 1)$-squares are defined recursively as follows:

(i) A 1-square is a cross at the odd-odd position,

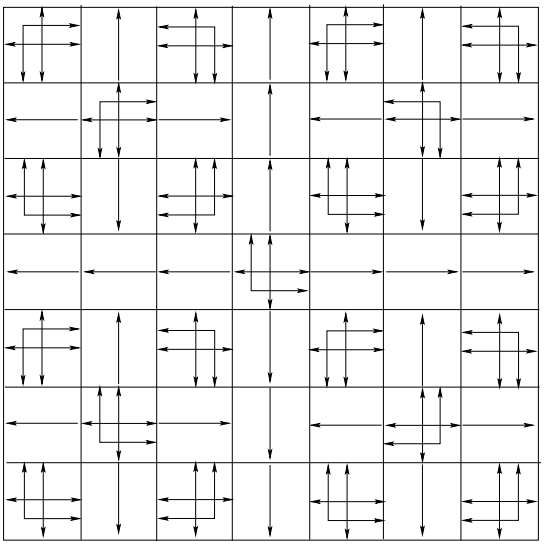
(ii) A $(2^{n+1} - 1)$-square consists of a cross in the middle (in an even-even position), sequences of arms radiating out of the center and four copies of $(2^n - 1)$-squares facing each other at the four quadrants:



Note that for every n there are actually four different $(2^n - 1)$-squares as the cross at the center may be in any of the four possible orientations. For example, the following figure illustrates the 3-square facing north and east:

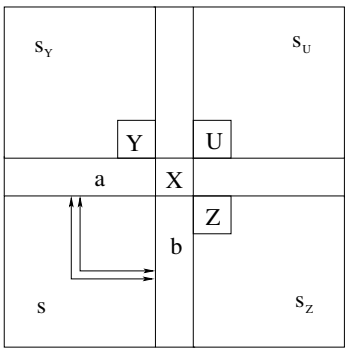


and the following figure shows the 7-square facing north and east. (For clarity, only the central, principal arrows of the arms are shown. The other arrows are uniquely determined by the orientations of the crosses.)



Inductively one easily gets the following properties of $(2^n - 1)$-squares: (1) The tiling is valid within the square, (2) all edges on the border of the square have arrow heads pointing out of the square, so all edge neighbors of $(2^n - 1)$-squares are forced to be arms, and (3) the only side arrows on the border are in the middle of the borders in the directions where the center cross of the square faces.

Consider an arbitrary valid tiling of the plane by Robinson's tiles. Let us show, using mathematical induction on n, that every cross in odd-odd position belongs to a unique $(2^n - 1)$-square, for every $n = 1, 2, \ldots$. The case $n = 1$ is trivial, as by definition 1-squares are themselves the crosses at odd-odd positions. Suppose then the claim is true for n and let C be an arbitrary cross in an odd-odd position. By the inductive hypothesis C belongs to a unique $(2^n - 1)$-square s. There are four possibilities for the orientation of this square, but they are all symmetric. Let us assume without loss of generality that s faces north and east. In the following discussion we refer to symbols indicating positions in the following figure:



First we prove that tile X, outside the north-east corner of square s, must be a cross. Suppose the opposite: X is an arm. Then it has an incoming arrow on all but one side, so one of its edge neighbors in regions a or b must be an arm directed towards X. By continuing this reasoning we see that all tiles in one of the regions a or b must be arms directed towards X. But this means that the tile at the center of region a or b is an arm with an incoming side arrow at the wrong end of the principal arrow: the side arrows are only possible towards the head of the principal arrow. Hence the assumption that X is an arm must be incorrect, and X must be a cross.

Consider then tile Y that is a cornerwise neighbor of X. It is in an odd-odd position and therefore Y is a cross. According to the inductive hypothesis Y belongs to a $(2^n - 1)$-square s_Y. This square cannot overlap with square s because then the tiles in the overlap region would belong to two different $(2^n - 1)$-squares which contradicts the uniqueness property. Also the tile north of X cannot belong to s_Y because X is a cross. Hence Y has to be at the south-east corner of s_Y. Analogously, tiles Z and U are corners of disjoint $(2^n - 1)$-squares s_Z and s_U, respectively. Tiles between these $(2^n - 1)$-squares are forced to be arms radiating out from X. The side arrows at the middle of a and b force the center crosses of s_Y and s_Z to face squares s and s_U, so the squares of s, s_Y, s_Z, s_U and the tiles between them form a $(2^{n+1} - 1)$-square that contains tile C.

We have proved the existence of a $(2^{n+1} - 1)$-square that contains C. The uniqueness is obvious as the orientation of the (unique) $(2^n - 1)$-square s that contains C determines the location of the center of the $(2^{n+1} - 1)$-square that contains C.

We have proved that every 1-square belongs to a 3-square, which belongs to a 7-square, which belongs to a 15-square and so on. Based on this observation we can state:

Lemma 4.4 *Robinson's tiles form an aperiodic protoset.*

Proof. The $(2^n - 1)$-squares are valid tilings of arbitrarily large squares, so a valid tiling of the plane exists (Corollary 4.3).

The centers of the quadrants of any $(2^n - 1)$-square are crosses separated by $(2^{n-1} - 1)$ arms. As every valid tiling contains $(2^n - 1)$-squares for every n, the tiling contains horizontally aligned crosses separated by arbitrarily long sequences of arms. So there can be no horizontal period, and a periodic tiling is not possible.

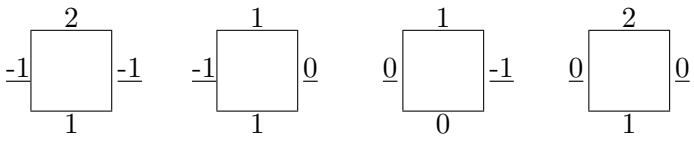
□

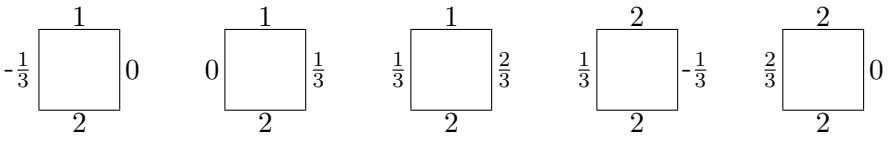
4.4 An aperiodic set of 14 Wang tiles

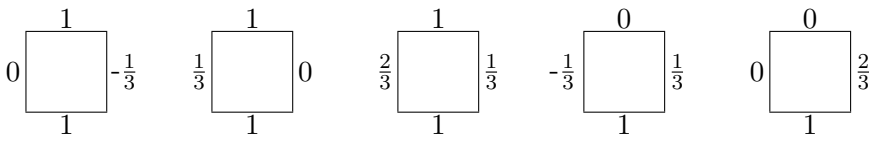
We have learned the aperiodic set of 56 Wang tiles by Robinson. In this section we learn a very different method of constructing aperiodic tile sets that yields a set with only 14 tiles, shown in the figure below. But note that even smaller aperiodic sets exist: E. Jeandel and M. Rao have an aperiodic Wang tile set that contains just 11 tiles, and they proved that 11 is the smallest possible size.

In our 14 tile set, the edges are labeled with rational numbers. Each number represents one color, so in valid tilings neighboring tiles must match in the numbers at the abutting edges. Notice also the the labels of the vertical edges of the first four tiles are underlined: This means that those numbers represent a different color than the same numbers without a line underneath.

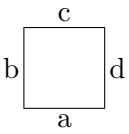
The set consists of two parts: the first four tiles form the set $\mathcal{P}_2$ and the set of the last ten tiles is called $\mathcal{P}_{2/3}$. The aperiodic set $\mathcal{P}$ is the union of these two sets. As the vertical sides of the elements of the two parts have different labels, it is clear that on any valid tiling of the plane by $\mathcal{P}$, each horizontal row is tiled by tiles that come from $\mathcal{P}_2$ or $\mathcal{P}_{2/3}$ only.







The tiles perform arithmetic operations in the following sense: We say that tile



multiplies by q if $qa + b = c + d$. In other words, the tile multiplies the "input" number a on its bottom edge by q, adds the "carry forward" b from the left edge, and splits the result between the "output" c at the top edge and the "carry forward" d to the right. It is easy to verify that the tiles in $\mathcal{P}_2$ all multiply by 2, and the tiles in $\mathcal{P}_{2/3}$ multiply by $\frac{2}{3}$.

Consider a horizontal segment of n tiles that all multiply by the same number q. Let a_i, b_i, c_i and d_i be the numbers on the i'th tile so that $qa_i + b_i = c_i + d_i$, for all $i = 1, 2, \ldots n$. Summing up over all n tiles we get

$$q\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i = \sum_{i=1}^{n} c_i + \sum_{i=1}^{n} d_i.$$

If the tiling constraint is satisfied then we have $d_i = b_{i+1}$ for all $i = 1, 2, \ldots, n-1$, and if the segment also starts and ends with the same carry forward $d_n = b_1$ we have that

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} b_i.$$

This happens if the segment is extracted from a periodic tiling with horizontal period n. Then

$$q\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} c_i.$$

Theorem 4.5 *The set $\mathcal{P}$ of 14 Wang prototiles above is aperiodic.*

Proof. We have two facts to prove: (i) no periodic tiling is possible, and (ii) some valid tiling exists.

(i) Suppose the opposite is true: there exists a periodic tiling $f : \mathbb{Z}^2 \longrightarrow \mathcal{P}$. Then we know that such a periodic tiling must have a horizontal period h and a vertical period v, for some $h, v > 0$. (In fact we could choose these two numbers to be identical.)

Let $a_{i,j}, b_{i,j}, c_{i,j}$ and $d_{i,j}$ be the colors on the south, west, north and east edges of the tile $f(i,j)$ in position $(i,j) \in \mathbb{Z}^2$. It follows from the tiling rule that $d_{i,j} = b_{i+1,j}$ and that $c_{i,j} = a_{i,j+1}$. As discussed above, we have

$$q_j \sum_{i=1}^{h} a_{i,j} = \sum_{i=1}^{h} c_{i,j} = \sum_{i=1}^{h} a_{i,j+1},$$

where $q_j = 2$ or $\frac{2}{3}$ depending on whether the tiles on row j come from set $\mathcal{P}_2$ or $\mathcal{P}_{2/3}$. By combining these equations for rows $j = 1, 2, \ldots, v$ we get the result that

$$q_1 q_2 q_3 \ldots q_v \sum_{i=1}^{h} a_{i,1} = \sum_{i=1}^{h} a_{i,v+1} = \sum_{i=1}^{h} a_{i,1}.$$

It is clear from the tiles that we cannot have a horizontal row of tiles such that the bottom edges all have value 0, so we have $\sum_{i=1}^{h} a_{i,1} > 0$. Hence we can divide $\sum_{i=1}^{h} a_{i,1}$ from the equation, which leaves

$$q_1 q_2 q_3 \ldots q_v = 1.$$

But each q_j is either 2 or $\frac{2}{3}$, and any product of these numbers is some power of 2 divided by a power of 3. Numbers 2 and 3 are relative primes, so no such product can equal 1, a contradiction.

(ii) It is enough to construct one valid tiling. The tiling will of course be non-periodic. We use the following notations and concepts. For any real number x, the floor $\lfloor x \rfloor$ of x is the largest integer not greater than x, that is, $\lfloor x \rfloor$ is the unique integer that satisfies $x - 1 < \lfloor x \rfloor \leq x$. Analogously, the ceiling $\lceil x \rceil$ is the smallest integer that is not smaller than x. By the *balanced representation* $B(x)$ of real number x we mean the bi-infinite sequence $\ldots B(x)_{-1}, B(x)_0, B(x)_1, B(x)_2, \ldots$ whose i'th term is $B(x)_i = \lfloor ix \rfloor - \lfloor (i-1)x \rfloor$. Notice that the elements of the sequence are integers, and

$$\begin{aligned} B(x)_i &= \lfloor ix \rfloor - \lfloor (i-1)x \rfloor < ix - ((i-1)x - 1) = x + 1, \text{ and} \\ B(x)_i &= \lfloor ix \rfloor - \lfloor (i-1)x \rfloor > ix - 1 - (i-1)x = x - 1. \end{aligned}$$

Hence each element of the sequence $B(x)$ is either $\lfloor x \rfloor$ or $\lceil x \rceil$.

Consider an arbitrary real number $x \in \left[\frac{1}{2}, 1\right]$. We have that $0 \leq x \leq 1$ and $1 \leq 2x \leq 2$. The symbols in the balanced sequences for x and $2x$ are 0's and 1's, and 1's and 2's, respectively. Let us show that the prototiles of $\mathcal{P}_2$ admit a tiling of an bi-infinite horizontal strip whose bottom labels read the sequence $B(x)$ and the top labels read $B(2x)$. Let

$$\begin{aligned} a_i &= B(x)_i, \\ b_i &= 2\lfloor (i-1)x \rfloor - \lfloor (i-1)(2x) \rfloor, \\ c_i &= B(2x)_i, \text{ and} \\ d_i &= 2\lfloor ix \rfloor - \lfloor i(2x) \rfloor \end{aligned}$$

be the labels on the south, west, north and east edges of the tile in position $i \in \mathbb{Z}$ of the strip. It is clear from this definition that $b_i = d_{i-1}$ so the labels match on the tiling of the strip. Let us analyze values of a_i, b_i, c_i and d_i to prove that the tile with these labels is in our set $\mathcal{P}_2$.

From the properties of the balanced sequences we know that $a_i \in \{0, 1\}$ and $c_i \in \{1, 2\}$. Clearly,

$$\begin{aligned} d_i - b_i &= 2\lfloor ix \rfloor - \lfloor i(2x) \rfloor - (2\lfloor (i-1)x \rfloor - \lfloor (i-1)(2x) \rfloor) \\ &= 2(\lfloor ix \rfloor - \lfloor (i-1)x \rfloor) - (\lfloor i(2x) \rfloor - \lfloor (i-1)(2x) \rfloor) \\ &= 2a_i - c_i, \end{aligned}$$

so the tiles of the strip multiply by number 2. We also have

$$d_i = 2\lfloor ix \rfloor - \lfloor i(2x) \rfloor < 2ix - (2ix - 1) = 1,$$

and

$$d_i = 2\lfloor ix \rfloor - \lfloor i(2x) \rfloor > 2(ix-1) - 2ix = -2.$$

Because d_i is an integer, the only possible values of d_i (and hence also b_i) are -1 and 0. The following possibilities remain:

$$\begin{array}{lcll}
a_i = 0, c_i = 2 & \Longrightarrow & d_i - b_i = 2a_i - c_i = -2, & \text{not possible,} \\
a_i = 0, c_i = 1 & \Longrightarrow & d_i - b_i = 2a_i - c_i = -1 & \Longrightarrow \ d_i = -1, b_i = 0, \\
a_i = 1, c_i = 2 & \Longrightarrow & d_i - b_i = 2a_i - c_i = 0 & \Longrightarrow \ d_i = b_i = -1 \text{ or } d_i = b_i = 0, \\
a_i = 1, c_i = 1 & \Longrightarrow & d_i - b_i = 2a_i - c_i = 1 & \Longrightarrow \ d_i = 0, b_i = -1.
\end{array}$$

Only four possibilities exist, and these are precisely the four tiles in $\mathcal{P}_2$.

Next we analyze $\mathcal{P}_{2/3}$ in a similar way. Let $x \in [1,2]$, so that $1 \le x \le 2$ and $\frac{2}{3} \le \frac{2}{3}x \le \frac{4}{3}$. The balanced representations of x and $\frac{2}{3}x$ consist of 1's and 2's, and 0's, 1's and 2's, respectively. Let us show that there is a tiling by $\mathcal{P}_{2/3}$ of a bi-infinite strip such that the labels on the bottom and the top of the strip read the balanced representations $B(x)$ of x and $B(\frac{2}{3}x)$ of $\frac{2}{3}x$. For brevity, let us denote $q = \frac{2}{3}$. The tile in position i of the strip has labels

$$\begin{array}{rcl}
a_i & = & B(x)_i, \\
b_i & = & q\lfloor (i-1)x \rfloor - \lfloor (i-1)(qx) \rfloor, \\
c_i & = & B(qx)_i, \text{ and} \\
d_i & = & q\lfloor ix \rfloor - \lfloor i(qx) \rfloor.
\end{array}$$

The consecutive tiles of the strip match as $b_i = d_{i-1}$. We know that $a_i \in \{1,2\}$ and $c_i \in \{0,1,2\}$. As above, we also have

$$\begin{array}{rcl}
d_i - b_i & = & q\lfloor ix \rfloor - \lfloor i(qx) \rfloor - (q\lfloor (i-1)x \rfloor - \lfloor (i-1)(qx) \rfloor) \\
& = & q(\lfloor ix \rfloor - \lfloor (i-1)x \rfloor) - (\lfloor i(qx) \rfloor - \lfloor (i-1)(qx) \rfloor) \\
& = & qa_i - c_i.
\end{array}$$

We also have

$$d_i = q\lfloor ix \rfloor - \lfloor i(qx) \rfloor < qix - (qix - 1) = 1,$$

and

$$d_i = q\lfloor ix \rfloor - \lfloor i(qx) \rfloor > q(ix-1) - qix = -q.$$

Because d_i is an integer multiple of $\frac{1}{3}$, the only possible values of d_i (and hence also b_i) are $-\frac{1}{3}, 0, \frac{1}{3}$ and $\frac{2}{3}$. The following possibilities remain:

$$\begin{array}{lcl}
a_i = 1, c_i = 2 & \Longrightarrow & d_i - b_i = qa_i - c_i = -\frac{4}{3}, \\
& & \text{not possible,} \\
\\
a_i = 1, c_i = 1 & \Longrightarrow & d_i - b_i = qa_i - c_i = -\frac{1}{3} \\
& \Longrightarrow & d_i = -\frac{1}{3}, b_i = 0 \text{ or } d_i = 0, b_i = \frac{1}{3} \text{ or } d_i = \frac{1}{3}, b_i = \frac{2}{3}, \\
\\
a_i = 1, c_i = 0 & \Longrightarrow & d_i - b_i = qa_i - c_i = \frac{2}{3} \\
& \Longrightarrow & d_i = \frac{1}{3}, b_i = -\frac{1}{3} \text{ or } d_i = \frac{2}{3}, b_i = 0, \\
\\
a_i = 2, c_i = 2 & \Longrightarrow & d_i - b_i = qa_i - c_i = -\frac{2}{3} \\
& \Longrightarrow & d_i = -\frac{1}{3}, b_i = \frac{1}{3} \text{ or } d_i = 0, b_i = \frac{2}{3}, \\
\\
a_i = 2, c_i = 1 & \Longrightarrow & d_i - b_i = qa_i - c_i = \frac{1}{3} \\
& \Longrightarrow & d_i = 0, b_i = -\frac{1}{3} \text{ or } d_i = \frac{1}{3}, b_i = 0 \text{ or } d_i = \frac{2}{3}, b_i = \frac{1}{3}, \\
\\
a_i = 2, c_i = 0 & \Longrightarrow & d_i - b_i = qa_i - c_i = \frac{4}{3}, \\
& & \text{not possible.}
\end{array}$$

The ten possibilities are exactly the tiles of $\mathcal{P}_{2/3}$.

We have proved that for every $x \in \left[\frac{1}{2}, 1\right]$ we can tile a strip whose bottom and top edges read the sequences $B(x)$ and $B(2x)$, respectively, and for every $x \in [1, 2]$ we can tile a strip whose bottom and top edges read the sequences $B(x)$ and $B(\frac{2}{3}x)$. Define the following real function $f : \left(\frac{2}{3}, 2\right] \longrightarrow \left(\frac{2}{3}, 2\right]$:

$$f(x) = \begin{cases} 2x, & \text{if } x \leq 1, \text{ and} \\ \frac{2}{3}x, & \text{if } x > 1. \end{cases}$$

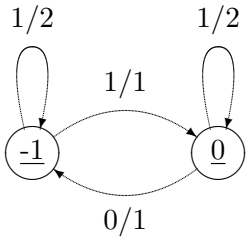
It is easy to see that the range of f is the half-open interval $\left(\frac{2}{3}, 2\right]$, so function f is surjective. (In fact, function f is a bijection, but this fact is not relevant to the reasoning below.) It follows from the surjectivity of f that there exist bi-infinite sequences $\ldots x_{-1}, x_0, x_1, x_2, \ldots$ of real numbers such that $x_{j+1} = f(x_j)$ for all $j \in \mathbb{Z}$. In fact, since one element x_0 of the sequence can be chosen arbitrarily from the half-open interval $\left(\frac{2}{3}, 2\right]$, the number of such sequences is uncountably infinite.

As proved above, for each $j \in \mathbb{Z}$ we can tile an infinite strip whose edges read $B(x_j)$ and $B(x_{j+1})$. By stacking these strips on top of each other we obtain a tiling of the plane. In fact, we proved there exist uncountably many different valid tilings.

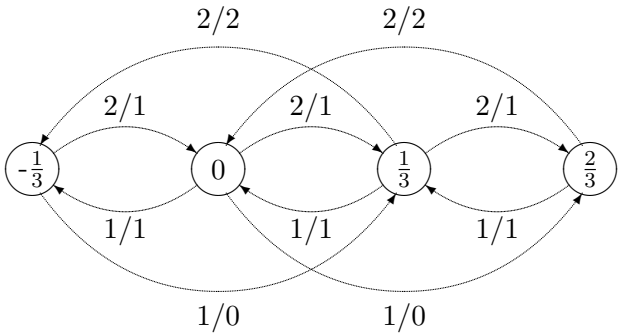
□

The following diagram illustrates the tiles as a directed graph whose nodes are labeled by the vertical colors, and the edges are labeled by pairs of "input"/"output" symbols. Each edge corresponds to a tile: the tile with labels a, b, c and d is the edge from node b to node d that is labeled by a/c. Any bi-infinite path through the diagram that follows the edges gives a valid tiling of one bi-infinite strip. Such a diagram is called a finite state transducer.

M_2 :



$M_{2/3}$:



Analogously, it is easy to construct for any given rational number q a finite set of tiles that multiply balanced sequences representing numbers of a given interval by q. The requested tiles have edge labels

$$\begin{array}{rcl} a & = & B(x)_i, \\ b & = & q\lfloor (i-1)x \rfloor - \lfloor (i-1)(qx) \rfloor, \\ c & = & B(qx)_i, \text{ and} \\ d & = & q\lfloor ix \rfloor - \lfloor i(qx) \rfloor. \end{array}$$

where x is a number in the desired interval and $i \in \mathbb{Z}$. A simple analysis shows that we always have $-q < b < 1$ and $-q < d < 1$, so there only is a finite number of such tiles.

The balanced sequences we used in the proof have many interesting properties. Balanced sequences of rational numbers are periodic, but if x is irrational then $B(x)$ is non-periodic, but only "barely so": it is a so-called Sturmian sequence, which means that the number of different subsegments of length n is $n+1$, for every n. This is the smallest possible number of different subsegments of length n in any non-periodic infinite sequence.

5 Undecidable problems concerning tiles

The following question (known as the domino problem or the tiling problem) arises naturally: How can one determine if a given finite set of Wang prototiles admits a tiling ? Does there exist some simple (or even complicated) properties that one can use to develop a computer program to determine if a tiling is possible. The input to the program should be an arbitrary finite set of Wang tiles, and the output should be "yes" or "no" depending on whether the input admits a tiling. In this section we show that such a computer program does not exist. The non-existence is a mathematical fact that cannot be overcome by building more powerful computers or by developing new programming languages or tools. The undecidability will be deduced from Turing's result on the undecidability of the halting problem of Turing machines, using the reduction technique.

The tiling problem is an example of a decision problem. A decision problem is a problem that has an input parameter, and the answer to the problem is always "yes" or "no". When we fix the value of the input parameter we get an instance of the problem. An instance is called a "yes"-instance or a "no"-instance depending on whether the answer to the decision problem is "yes" or "no", respectively. For example, the problem "Does a given quadratic polynomial have a real root ?" is a decision problem. Quadratic polynomials are instances (we always use the keyword "given" in the decision problem statement to indicate the input). For example, x^2-2x+2 is a "no"-instance, while x^2+2x-1 is a "yes"-instance to this problem. In the tiling problem, instances are finite sets of Wang tiles, and an instance is a "yes"-instance iff the protoset admits a tiling of the plane. The complement of a decision problem is the decision problem where the "yes" and the "no" answers have been switched: for example, the complement of the tiling question asks whether the given protoset does not admit a tiling. "Yes" means now that no tiling is possible.

An algorithm can be formally defined in various ways. In order to keep the discussion simple, we are going to leave it undefined. For our purposes it is sufficient to understand a (decision) algorithm to be a computer program that takes some input and returns a "yes" or a "no" answer on each input. We say that the algorithm solves a decision problem if the algorithm returns the correct yes/no -answer on every instance of the problem. If such an algorithm exists then the decision problem is called decidable and if no such algorithm exists then the problem is undecidable.

Strictly speaking, the input of a computer program is a string of bits (or a string over some other alphabet) so the instance of a decision problem has to be encoded into such a string. For example, Wang tiles could be encoded as sequences of four colors, each color represented as a unique binary string. There are of course many ways to do such an encoding, but all encodings are equivalent in the sense that the decidability status of the decision problem is not affected by the encoding. In our discussion encodings of inputs will be irrelevant as we are not going to write any actual programs – rather algorithms will be defined by describing in plain English the steps that the algorithm executes. The idea is that we all are sufficiently familiar with computer programming so that such a description (when detailed enough) will convince everyone that a computer program exists for solving the program. Notice also that when describing an algorithm we do not need to worry about things related to computing resources such as memory space etc. We are always supposed to have unlimited amounts of such resources.

For example, the following algorithm solves the question whether a given quadratic polynomial has a real root: Let ax^2+bx+c be the input to the algorithm. The algorithm starts by computing the discriminant $D=b^2-4ac$. Then it checks whether $D \geq 0$ or not. If $D \geq 0$ then the algorithm returns

"yes", and if $D < 0$ then the algorithm returns "no". This level of description should convince everyone that the problem is decidable.

Notice that from the decidability point of view any problem that has no input instance or has only a finite number of possible instances is trivially decidable. For example, a question like "Is the Riemann hypothesis true ?" is decidable. There is a trivial algorithm that solves this problem, we just do not know what that algorithm is. (The algorithm is either the algorithm that only types "yes" or the algorithm that only types "no".) In the same way, any problem with a finite number of possible input instances is solved by a program that simply looks-up from a finite table the correct answer corresponding to the given input. For example, for any fixed number N it is decidable if a given set of N Wang tiles admits a valid tiling, as there are only finitely many such sets, up to renaming of the colors. So decidability questions are only relevant in connection to decision problems with an unlimited number of possible input instances.

A semi-algorithm is a weaker concept than an algorithm: it is a computer program that halts and returns "yes" if the input is a positive instance of the problem, but it may run forever, without ever halting, on negative input instances. So a semi-algorithm (semi-) solves a decision problem if on every "yes" -instance of the problem it returns the correct "yes" -answer, but when the input is a "no" -instance it may run forever without ever returning an answer. If an answer is returned, it has to be the correct answer: semi-algorithms never return wrong answers. If a decision problem has a semi-algorithm then it is called semi-decidable. Clearly every decidable problem is also semi-decidable as an algorithm is also a semi-algorithm. As an example of a semi-algorithm that is not an algorithm consider the following process of determining if a given Wang protoset does not admit a tiling.

Lemma 5.1 *The complement of the tiling problem "Does a given finite set $\mathcal{P}$ of Wang prototiles admit a tiling ?" is semi-decidable.*

Proof. The semi-algorithm enumerates positive integers $n = 1, 2, 3, \ldots$ one-by-one. For each n it tries all possible ways of tiling the $n \times n$ square by the given prototiles. It can simply try (in the lexicographic order) each sequence of n^2 tiles, write the tiles inside the $n \times n$ square row-by-row and check whether the tiles match or not. For each n there are only a finite number of sequences to try, and an algorithm can easily go through all of them one-by-one. If we find a valid tiling of the $n \times n$ square we increment n and repeat. If we do not find a tiling of the $n \times n$ square then a valid tiling of the plane does not exist, so the semi-algorithm returns answer "yes" to indicate that no tiling is admitted. Notice that if the given instance is a "no"-instance (i.e. admits a tiling) the process will never end as we keep on tiling larger and larger squares. But on every "yes" -instance (i.e. no tiling exist) the process will terminate with the correct output, because by Corollary 4.3 some $n \times n$ square cannot be tiled. We conclude that the complement of the tiling problem is semi-decidable. □

We make the following observations:

Theorem 5.2 *A decision problem is decidable if and only if the complement problem is decidable. A decision problem is decidable iff the problem and its complement are both semi-decidable.*

Proof. If decision problem P is decidable then there exists an algorithm A that solves P. We get an algorithm for the complement problem if we simply switch the output of A, or more precisely, make a new algorithm A' that (i) calls A as a subroutine with the original input of A', and (ii) if A returns "yes" algorithm A' returns "no" and if A returns "no" then A' returns "yes". This proves the first claim.

Algorithm is also a semi-algorithm so if P is decidable it is also semi-decidable and, since the complement problem is decidable, the complement problem is also semi-decidable. Conversely, suppose that problem P has a semi-algorithm A and the complement of P has a semi-algorithm A'. Here is a description of an algorithm that solves P: With a given input we execute both A and A' at the same time. This

can be arranged by, for example, alternating between the semi-algorithms by executing one step of each semi-algorithm in turn. Eventually one of the two semi-algorithms will return the "yes"-answer. If the instance is a "yes" -instance then it will be A that gives the answer and if the instance is a "no" -instance then A' will give the answer. In the first case our algorithm returns "yes", in the second case we return "no".

□

In view of the previous theorem and lemma: if the tiling problem were semi-decidable then it would be also decidable (because we know that the complement problem is semi-decidable). Equivalently: Once we prove that the tiling problem is undecidable, it implies that the tiling problem is not semi-decidable either.

In order to prove from the scratch that a problem is undecidable we would need a more precise definition of an algorithm. Here, to avoid too deep and time consuming involvement in the computation theory (which a topic of another course) we are not going to prove undecidability results from the scratch. Rather, we take the classic result by Alan Turing without a proof. This result provides us with one decision problem (the halting problem of Turing machines without input) that is known to be undecidable. Turing's proof of this result is a very nice — and not very difficult — diagonal argument similar to Cantor's proof for the uncountability of real numbers.

Once we have one undecidable decision problem available, we use the technique of reduction to prove other problems undecidable. Reduction is an indirect proof technique that works as follows: Suppose P is a known undecidable problem (e.g. the halting problem of Turing machines), and we want to prove that problem Q is also undecidable. We make the assumption (indirect proof!) that there exists an algorithm A that solves Q. Then we describe an algorithm that solves P, using A as a subroutine. Since P is undecidable, no such algorithm can exist, so algorithm A cannot exist either. In the reduction technique we design an algorithm for P in order to prove that no algorithm exists that solves Q.

5.1 Turing machines

Let us start by defining Turing machines. A Turing machine is a simple computing device that consists of a bi-infinite tape that serves as the memory and a finite state processor that moves over the tape. The tape consists of a sequence of memory locations, indexed by $\mathbb{Z}$, each of which contains an element of a finite set Γ, called the tape alphabet. So the content of the tape at any given time is given by a function $f : \mathbb{Z} \longrightarrow \Gamma$ where $f(i)$ is the symbol at location i. One element $b \in \Gamma$ is specified as the blank symbol, and in the beginning of the computation all tape locations contain symbol b.

At all times, the processor (also called the control unit) of the machine accesses one tape location $i \in \mathbb{Z}$. The control unit is in some state q that is an element of a finite state set S. Depending on the current state q and the current tape symbol $f(i) \in \Gamma$ at the current location i of the control unit, the Turing machine changes the state of the control unit, replaces the tape symbol $f(i)$ with a new symbol, and moves the control unit one position to the left or right on the tape. This action of the machine is specified by its transition function

$$\delta : S \times \Gamma \longrightarrow S \times \Gamma \times \{L, R\}.$$

The interpretation of $\delta(q, x) = (p, y, d)$ is that if the current state is q and the tape symbol at the current location i is x then the machine changes the state into p, replaces x by y on the tape and moves one position left or right on the tape depending on whether $d = L$ or $d = R$.

The configuration of the machine is an element of $S \times \mathbb{Z} \times \Gamma^{\mathbb{Z}}$. Configuration (q, i, f) specifies the current state q of the machine, its current position i on the tape, and the current content f of the entire tape. Formally we can now define one move of the machine: Configuration (q, i, f) is transformed in one move into the configuration (p, j, g), where $\delta(q, f(i)) = (p, y, d)$, $g(i) = y$, $g(k) = f(k)$ for all $k \neq i$, and

$j = i + 1$ if $d = R$ and $j = i - 1$ if $d = L$. We denote this move by

$$(q, i, f) \vdash (p, j, g).$$

In the beginning of the computation the Turing machine is in one specific state $s_0 \in S$ called the initial state, and another state $s_h \in S$ is specified as the halting state. The Turing machine halts when the control unit enters state s_h. The Turing machine then can be understood as a dynamical system where the transformation $\vdash$ is applied repeatedly starting from the initial configuration $(s_0, 0, f_b)$ where $f_b(k) = b$ for all $k \in \mathbb{Z}$ until (if ever) the machine reaches and halts in some configuration (s_h, i, f), where i and f can be arbitrary.

To specify a Turing machine one needs to provide six items. We say that a Turing machine is a six-tuple $M = (S, \Gamma, \delta, s_0, s_h, b)$ where S and Γ are finite sets, $s_0, s_h \in S$ and $b \in \Gamma$ are elements of those sets, and $\delta : S \times \Gamma \longrightarrow S \times \Gamma \times \{L, R\}$ is a function. Note that in the most common terminology in the literature one also specifies a third finite set, the input alphabet, but we can ignore it here because we only discuss Turing machines without input.

Example 9. Consider the following Turing machine $M = (\{s, t, h\}, \{a, b\}, \delta, s, h, b)$ where

$$\begin{array}{rcl} \delta(s, a) & = & (t, a, L) \\ \delta(s, b) & = & (t, a, R) \\ \delta(t, a) & = & (h, a, L) \\ \delta(t, b) & = & (s, a, L) \end{array}$$

(and the values of $\delta(h, \dots)$ do not matter as h is the halting state.) The computation by M proceeds as follows:

$$\dots bbbb \overset{s}{b} bb \cdots \vdash \dots bbbba \overset{t}{b} b \cdots \vdash \dots bbb \overset{s}{a} abb \cdots \vdash \dots bb \overset{t}{b} aab \cdots \vdash$$
$$\dots b \overset{s}{b} aaab \cdots \vdash \dots ba \overset{t}{a} aab \cdots \vdash \dots b \overset{h}{a} aaab \dots$$

□

Note that at all times only a finite number of tape locations may contain symbols that are different from the blank symbol b. Hence configurations (s, i, f) have a finite representation. The following result is given without proof:

Theorem 5.3 (Turing 1936) *There is no algorithm to solve the following decision problem: "Does a given Turing machine M eventually halt ?"*

□

Note that the decision problem of the previous theorem is semi-decidable: A simple semi-algorithm for the "yes" instances simply simulates the Turing machine step-by-step until it halts, if ever. Once the halting state q_h is reached, the semi-algorithm returns answer "yes". The "no" answer is never returned: if the Turing machine does not halt then the simulation continues indefinitely.

Based on Theorem 5.3 we can now prove other problems undecidable using the reduction technique discussed earlier.

Example 10. As an example of a reduction, let us prove that the following decision problem is undecidable: "Given a Turing machine M and a tape symbol a, does the Turing machine M eventually write symbol a somewhere on the tape ?" Let us call this problem Q, and let us call the halting problem of Theorem 5.3 problem P. Suppose we have an algorithm A that solves Q. Then there exists the following algorithm B to solve P:

Algorithm B gets as input a Turing machine $M = (S, \Gamma, \delta, q_0, q_h, b)$. In order to determine if M halts, algorithm B creates a new Turing machine $M' = (S \cup \{q'\}, \Gamma \cup \{a\}, \delta', q_0, q', b)$ where $q' \notin S$ is the new

halting state, $a \notin \Gamma$ is a new tape symbol, and δ' is exactly like δ, except for the following modification: For every $x \in \Gamma$ we set $\delta(q_h, x) = (q', a, R)$. For all $x \in \Gamma \cup \{a\}$ and $q \in S \cup \{q'\}$ the values of the new entries $\delta(q', x)$ and $\delta(q, a)$ of δ can be chosen arbitrarily. The idea is that the modified M' works exactly like M until (if ever) M enters the halting state q_h. Instead of halting in state q_h the new machine writes then the new tape letter a on the tape and halts. Clearly, M halts if and only if M' writes symbol a on the tape.

Algorithm B can easily construct M'. Then B gives this M' and symbol a as the input to the algorithm A. Algorithm A returns "yes" or "no" depending on whether M' eventually writes a on the tape or not. Algorithm B then simply returns that same answer.

We described an algorithm B that according to Theorem 5.3 does not exist. Therefore the assumption that algorithm A exists must be incorrect.

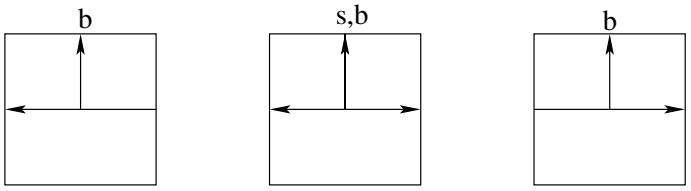
□

5.2 The tiling problem with a seed tile

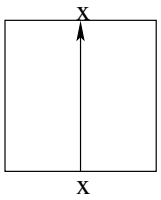
To use the reduction technique in connection to tiling questions we start by designing Wang protosets that simulate Turing machines. A valid tiling will picture the entire computation history by a Turing machine, move-by-move. Instead of colors on the edges of the tiles we use labeled arrows. In a valid tiling, each arrow head and tail must meet a tail and head, respectively, with the same label. The tiling constraints using such arrows can then be easily transformed into color constraints by replacing an arrow with label L and direction D by color (L, D), where D can be North, East, South or West.

The labels of the arrows will be tape symbols (representing a tape location containing that symbol) and state/tape symbol pairs (representing a tape location containing the control unit at the given state). Any given Turing machine $M = (S, \Gamma, \delta, s, h, b)$ will be represented by a set $\mathcal{P}_M$ of Wang tiles that contains

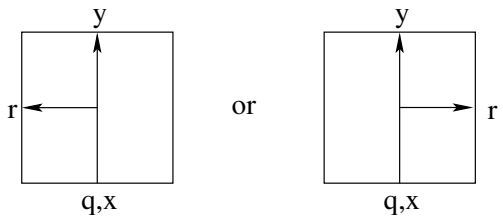
(i) the following three starting tiles to represent the blank tape



(ii) for every tape letter $x \in \Gamma$ an alphabet tile

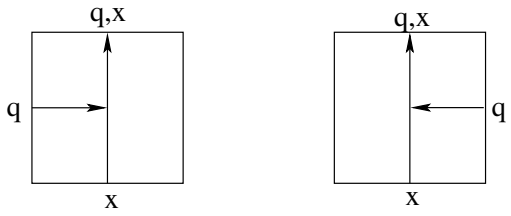


(iii) for every non-halting state $q \in S \setminus \{h\}$ and tape symbol $x \in \Gamma$ one action tile

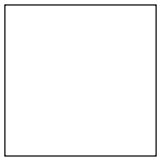


where the left tile is used iff $\delta(q, x) = (r, y, L)$ and the right tile iff $\delta(q, x) = (r, y, R)$,

(iv) for every non-halting state $q \in S \setminus \{h\}$ and tape symbol $x \in \Gamma$ the two merging tiles

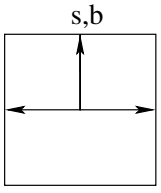


(v) the blank tile



Theorem 5.4 *The following decision problem is undecidable: "Given a finite set $\mathcal{P}$ of Wang prototiles and one specified seed tile $t \in \mathcal{P}$, does $\mathcal{P}$ admit a valid tiling of the plane that contains at least one occurrence of t ?"*

Proof. Suppose an algorithm exists. Then we can solve the halting problem of Turing machines as follows: For any given Turing machine M our algorithm starts by constructing the set $\mathcal{P}_M$ described above. This set can clearly be mechanically constructed based on M. Then set $\mathcal{P}_M$ and the seed tile



are given as input to the algorithm that determines if a tiling exists that contains the seed tile. Such a tiling exists if and only if M does not halt: The seed tile uniquely determines the entire tiling. Tiles on the same horizontal row as the seed must be starting tiles of type (i). Horizontal rows above are forced to represent consecutive configurations of the Turing machine. If the machine halts then the tiling becomes impossible as there are no action tiles of type (iii) for the halting state. But if the machine M never halts then the tiling can be continued indefinitely, to fill the entire upper half of the plane. The lower half plane can be filled with the blank tile of type (v).

□

Note that the decision problem in the previous theorem is not the same as the tiling problem. The request for the named seed tile to appear in the tiling makes the proof easy, as it allows us to guarantee the proper initialization of the Turing machine computation.

5.3 Finite systems of forbidden patterns

Before moving on to other decision problems concerning tiles, we simplify the discussion by relaxing the requirement to use colors or arrows to specify which tiles may be put next to each other. Rather, correctness of a tiling will be specified by a finite collection of *forbidden patterns*. A configuration is then a valid tiling if and only if it does not contain a forbidden pattern.

More precisely, let us first define a *neighborhood vector*

$$N = (\vec{n}_1, \vec{n}_2, \ldots, \vec{n}_m)$$

where each $\vec{n}_i \in \mathbb{Z}^2$ and $\vec{n}_i \neq \vec{n}_j$ for all $i \neq j$. The elements $\vec{n}_i$ specify the relative locations of the neighbors of each position: Position $\vec{n} \in \mathbb{Z}^2$ has m *neighbors* $\vec{n} + \vec{n}_i$ for $i = 1, 2, \ldots, m$.

Let T be a finite set, the set of prototiles, and let $R \subseteq T^m$ be a relation specifying which patterns are allowed in valid tilings: A configuration $c \in T^{\mathbb{Z}^2}$ is valid at position $\vec{n} \in \mathbb{Z}^2$ if and only if

$$[c(\vec{n} + \vec{n}_1), c(\vec{n} + \vec{n}_2), \ldots, c(\vec{n} + \vec{n}_m)] \in R,$$

that is, the neighborhood of $\vec{n}$ contains an allowed pattern. Configuration c is a valid tiling iff it is valid at at all positions $\vec{n} \in \mathbb{Z}^2$.

Note that the complement of R is the set of forbidden patterns: any configuration that contains such a pattern is not a valid tiling. The triplet (T, N, R) specifies valid tilings, and we call such a triplet a *finite system of forbidden patterns*.

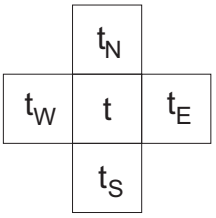
Any Wang tile set can be expressed as a finite system of forbidden patterns with the neighborhood

$$N = [(0, 0), (0, -1), (0, 1), (-1, 0), (1, 0)],$$

and the relation R that contains all those patterns where the colors of the tiles match:

$$(t, t_S, t_N, t_W, t_E) \in R$$

if and only if the colors match between t and the neighboring tiles in



There is also a correspondence to the other direction: for any finite system of forbidden patterns we can effectively (that is, algorithmically) construct a Wang tile set such that there is a natural correspondence between valid tilings in the two tile systems.

Lemma 5.5

(i) For every Wang protoset $\mathcal{P}$ one can effectively construct a finite system $S = (\mathcal{P}, N, R)$ of forbidden patterns over $\mathcal{P}$ such that $c \in \mathcal{P}^{\mathbb{Z}^2}$ is a valid Wang tiling if and only if c is valid according to S.

(ii) Conversely, for every finite system $S = (T, N, R)$ of forbidden patterns one can effectively construct a Wang protoset $\mathcal{P}$ such that $\mathcal{P}$ admits a (periodic) tiling if and only if S admits a (periodic) tiling.

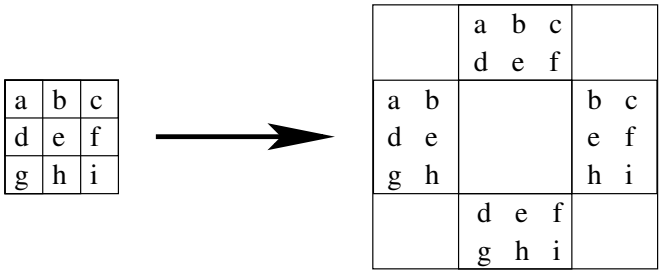
Proof. Part (i) is clear and was already explained above. Consider then (ii). Let $S = (T, N, R)$ be a finite system of forbidden patterns. Observe first that we can assume, without loss of generality, that elements of N form a square: Let $m \geq 2$ be an integer such that there is an $m \times m$ square M containing all elements $\vec{n}_i$ of the neighborhood vector N. We can add to N the missing elements of M – keeping them irrelevant in the relation R – and obtain a system that admits exactly same valid tilings and whose neighborhood vector consists exactly of the elements of M.

Assuming now that N forms an $m \times m$ square, we construct a set $\mathcal{P}$ of Wang tiles as follows: The tiles are the allowed $m \times m$ square patterns over T, that is, $\mathcal{P} = R$. Colors of $t \in R$ are obtained by erasing one boundary column or row from it: if $t = [t_{ij}]_{1 \leq i \leq m}^{1 \leq j \leq m}$ is a tile, where each $t_{ij} \in T$, then the left, top, right and bottom colors of t are

$$[t_{ij}]_{1 \leq i \leq m-1}^{1 \leq j \leq m}, \qquad [t_{ij}]_{1 \leq i \leq m}^{2 \leq j \leq m}, \qquad [t_{ij}]_{2 \leq i \leq m}^{1 \leq j \leq m}, \qquad [t_{ij}]_{1 \leq i \leq m}^{1 \leq j \leq m-1},$$

respectively. So the vertical colors are $(m-1) \times m$ blocks and horizontal colors are $m \times (m-1)$ blocks over T.

For example, the following figure illustrates the tile corresponding to a 3×3 allowed pattern:



Allowed 3x3 pattern Wang tile

Note that two adjacent Wang tiles then match if and only if the $m \times m$ patterns they represent have the correct $(m-1) \times m$ or $m \times (m-1)$ overlap when the tiles are placed next to each other.

This construction immediately implies that, if $f \in T^{\mathbb{Z}^2}$ does not contain any forbidden pattern, then the function $g \in \mathcal{P}^{\mathbb{Z}^2}$ is a correct Wang tiling, where for each $(i, j) \in \mathbb{Z}^2$ we set $g(i, j)$ be the $m \times m$ pattern in f whose lower left corner is in position (i, j).

Conversely, let $g \in \mathcal{P}^{\mathbb{Z}^2}$ be a valid Wang tiling. Let $f \in T^{\mathbb{Z}^2}$ be the function such that $f(i, j)$ is the symbol at the lower left corner of the $m \times m$ square $g(i, j)$, for every $(i, j) \in \mathbb{Z}^2$. Because of the overlap property between neighboring Wang tiles, we have that the $m \times m$ blocks extracted from f coincide with the corresponding tiles in g. Because the elements of $\mathcal{P}$ are the allowed $m \times m$ patterns we have that f only contains allowed $m \times m$ blocks, and therefore f is a valid tiling.

Note that in the previous reasoning f is periodic if and only if g is periodic.

□

The previous lemma means that in the following decision problems we can describe tiles in any terms that locally determine which tiles are allowed to be next to each other. Such tiles can anyway be effectively converted into an equivalent set of Wang tiles. This substantially simplifies the discussion.

5.4 The periodic tiling problem

Next we consider the problem of deciding if a given protoset admits a periodic tiling. There is an obvious semi-algorithm:

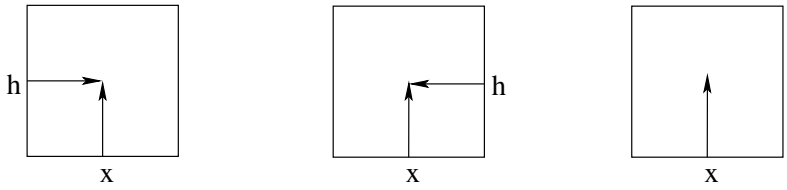
Lemma 5.6 *The decision problem "Does a given finite set $\mathcal{P}$ of Wang prototiles admit a periodic tiling ?" is semi-decidable.*

Proof. The semi-algorithm enumerates positive integers $n = 1, 2, 3, \ldots$ one-by-one, and for each n it constructs all valid tilings of the $n \times n$ square. There are only a finite number of them. For each such tiling, we check if the top and the bottom of the square read the same colors, and if the left and the right side also read the same colors. If they do, we have found a square pattern that can be repeated to form a periodic tiling of the plane. The semi-algorithm returns then "yes". We know that if a periodic tiling exists then some $n \times n$ square forms a period of a periodic tiling, so our semi-algorithm is guaranteed to correctly detect all "yes"-instances. □

Note that if no aperiodic tile sets existed then there would be an algorithm for determining if a periodic tiling exists: We namely have semi-algorithms for detecting that no tiling exists (Lemma 5.1) and that a periodic tiling exists (Lemma 5.6). If these were the only two possibilities then the two semi-algorithms would yield an algorithm (Theorem 5.2) that would determine if a (periodic) tiling exists. However, since aperiodic protosets do exist, this reasoning can not be used. In fact, the admittance of periodic tilings turns out to be undecidable. Not surprisingly, an aperiodic protoset is needed in the proof.

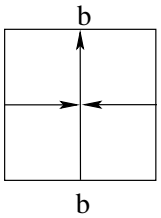
Theorem 5.7 *The following decision problem is semi-decidable but not decidable: "Does a given finite set $\mathcal{P}$ of Wang prototiles admit a periodic tiling ?"*

Proof. Semi-decidability was discussed above in Lemma 5.6. Let us prove undecidability via a reduction from the halting problem of Turing machines. For any given Turing machine M, we construct the Wang set $\mathcal{P}_M$ from Theorem 5.4. We add to these the following halting tiles for every tape letter $x \in \Gamma$:



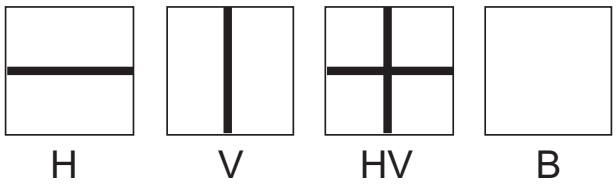
Here h is the halting state. The halting tiles have the effect that a tiling becomes possible even if the Turing machine halts — then the state component simply disappears from the configuration. Using the third tile, the entire configuration can then disappear.

We also add the following tile

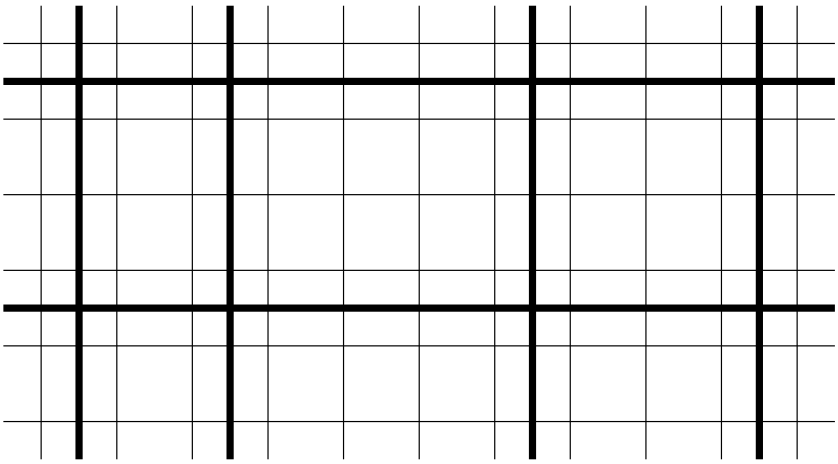


to the start tiles of the Turing machine. This tile allows the same horizontal row to contain several copies of the start configuration of the Turing machine.

In addition to $\mathcal{P}_M$, we take one aperiodic tile set $\mathcal{P}$. This can be, for example, the Robinson's aperiodic tile set. Finally, we also use the following set $\mathcal{Q}$ of prototiles: Set $\mathcal{Q}$ contains tiles with horizontal and/or vertical fault lines, as well as tiles without a fault line:



There are no tiles to end a fault line, so all fault lines cut the plane horizontally or vertically:



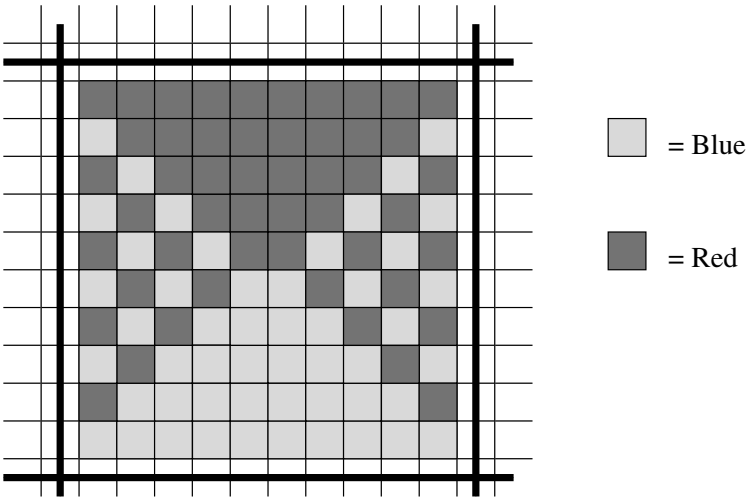
We want the fault line tiles to satisfy the following property:

(*) If a tiling contains at least two parallel fault lines then it also contains at least two fault lines in the perpendicular direction.

To establish this, we make two versions of the empty tile B without fault lines: one called red and the other one called blue. Then we add the following local constraints on validity of tilings:

- The northern neighbor of a horizontal fault line (tile H) must be a blue blank (=blue version of tile B).
- The southern neighbor of a horizontal fault line (tile H) must be a red blank (=red version of tile B).
- The northern neighbor of a blue blank whose horizontal neighbors are blue blanks is a blue blank.
- The southern neighbor of a red blank whose horizontal neighbors are red blanks is a red blank.

These local constraints are satisfied in tilings where the fault lines partition the plane into squares of even size. The insides of the squares consist of a blue and a red triangle and two triangles with the checker-board pattern of blue and red:



Notice that the constraints are local and can be implemented using a finite system of forbidden patterns.

Suppose that a tiling contains two horizontal fault lines at distance n from each other. Suppose there would be a horizontal segment of length $2n$ without a vertical fault line. Then there is a blue segment of length $2n$ on top of the lower fault line. On top of it we have $2n-2$ blue tiles, then $2n-4$ blue tiles, and so on. We see that the horizontal row below the upper fault line must contain blue tiles, which is not allowed by the local constraints above. Conclusion: if a tiling contains at least two horizontal fault lines then it also contains at least two vertical fault lines.

An analogous coloring is also done in the perpendicular direction. Hence the non-fault line tiles come if four varieties, one for each combination of blue/red color in horizontal/vertical direction. Then our set $\mathcal{Q}$ satisfies the required property (*).

After establishing the three tile sets $\mathcal{P}_M$ (simulates the Turing machine), $\mathcal{P}$ (aperiodic set) and $\mathcal{Q}$ (fault lines), the algorithm combines these three sets into a single tile set by taking their cartesian product

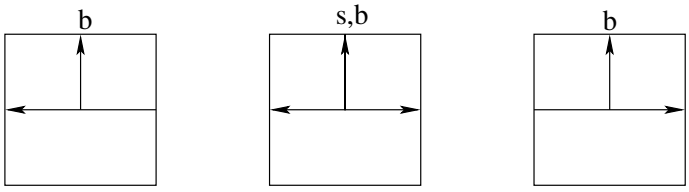
$$\mathcal{P}_M \times \mathcal{P} \times \mathcal{Q}.$$

Each of the three components of any tile must match locally with its neighbors according to the rules of the corresponding tile set. In this way tilings will be "sandwiches" with three layers. For any $(a,b,c) \in \mathcal{P}_M \times \mathcal{P} \times \mathcal{Q}$ we call a, b and c the first, the second and the third layer, respectively.

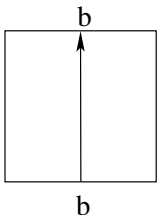
We add the following local constraints on tilings:

(1) In tile (a,b,c), if c contains a fault line then the tiling rule is not enforced on the second layer b. The idea is to allow periodic tilings (even though $\mathcal{P}$ is aperiodic) in the presence of fault lines.

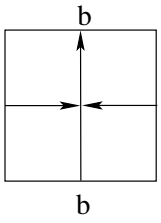
(2) In tile (a,b,c), if c contains only the horizontal fault line then the first component a must be one of the start tiles



(3) In tile (a,b,c), if c contains only the vertical fault line then the first component a must be



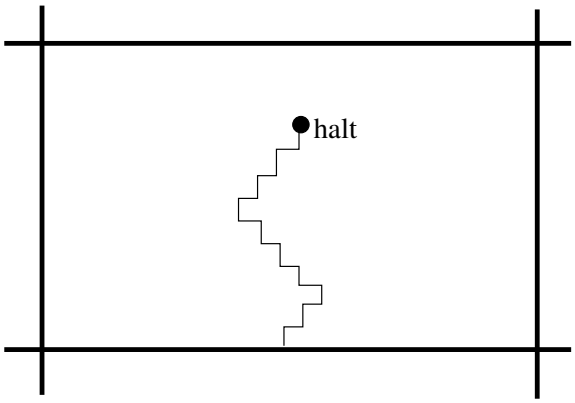
(4) In tile (a, b, c), if c contains both horizontal and vertical fault lines then a must be



These constraints force the lower border of any rectangle surrounded by fault lines to contain (a finite segment of) the blank tape and a single Turing machine in its initial state s. The vertical fault lines are forced to contain the black symbol only, and the Turing machine is never allowed to reach a vertical fault line.

The construction of the tile set is now complete. Suppose first that Turing machine M halts in n steps. Then the tiles admit a valid periodic tiling with the horizontal and vertical period $2n$. On the third layer the fault lines partition the space into squares of size $2n \times 2n$. The second layer contains a correctly tiled $2n \times 2n$ square, repeated inside the squares between the fault lines. The tiling of the second layer fails on some tiles along the fault lines, but that is allowed by (1) above.

The first layer consists of the halting simulation of the Turing machine M. The start of the simulation begins at the bottom of each $2n \times 2n$ square. The entire simulation fits inside the $2n \times 2n$ square, because the machine halts after n steps. The halting tiles allow the disappearance of the Turing machine configuration before the next vertical line is reached. Hence a periodic tiling is admitted.



Conversely, suppose a periodic tiling exists. Because $\mathcal{P}$ is an aperiodic set, there must be a place on the tiling where the tiling in the second layer is incorrect. This is possible only if there is a fault line in that location. Because the tiling is periodic, this implies the existence of infinitely many parallel fault lines. By property (*) this further implies the presence of a rectangle bordered by fault lines. Consider the first

layer of one such rectangle. The bottom is forced to contain a (finite segment) of the start configuration of the Turing machine. The tiles in $\mathcal{P}_M$ force the following rows to simulate the Turing machine moves one-by-one. If the Turing machine does not halt then the simulation continues without a limit and the Turing machine never disappears. But the next horizontal fault line is possible only if the Turing machine disappears. Hence the Turing machine must halt before the simulation reaches the upper border of the rectangle.

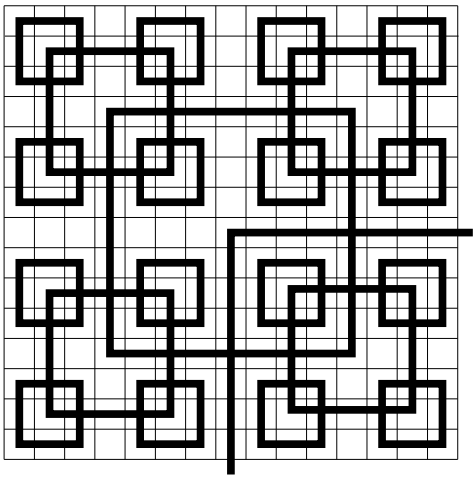
The tile set we constructed is given as input to our hypothetical algorithm that determines if the set admits a periodic tiling. (More precisely, the input is the corresponding set of Wang tiles, obtained by the conversion from the system of forbidden blocks as described in Lemma 5.5). We know that a periodic tiling is possible if and only if M halts, so we get the answer to the halting problem, a contradiction.

□

5.5 The tiling problem

Now we turn to the general tiling problem: "Does a given Wang tile set admit a valid tiling?" To prove undecidability we make a reduction from the tiling problem with a seed tile, proved undecidable in Theorem 5.4. Now we have no specified seed tile required to be used, so the main problem is how to force the presence of the seed (=the beginning state of the Turing machine) in every valid tiling. Note that if it is possible to have arbitrarily large squares without the seed, then it is also possible to make the entire tiling without the seed. This is a consequence of the compactness of the tiling space (Corollary 4.3). Therefore the seed must be enforced inside all $n \times n$ squares for some n. This on the other hand would seem to be contradictory to the possibility that tilings with only a single seed tile may occur. (Indeed, in our proof of Theorem 5.4 only a single seed tile occurs, starting an infinite, non-halting computation of a Turing machine.) A solution is to partition the space using Robinson's aperiodic tile set into "nested boards", each containing a copy of piece of a valid tiling around a seed tile.

This takes us back to the Robinson's tile set. Recall the special $(2^n - 1)$-squares that necessarily exist in every valid tiling. We define nested boards using the side arrows of Robinson's tiles. The following figure shows only the side arrows of a $(2^n - 1)$-square:



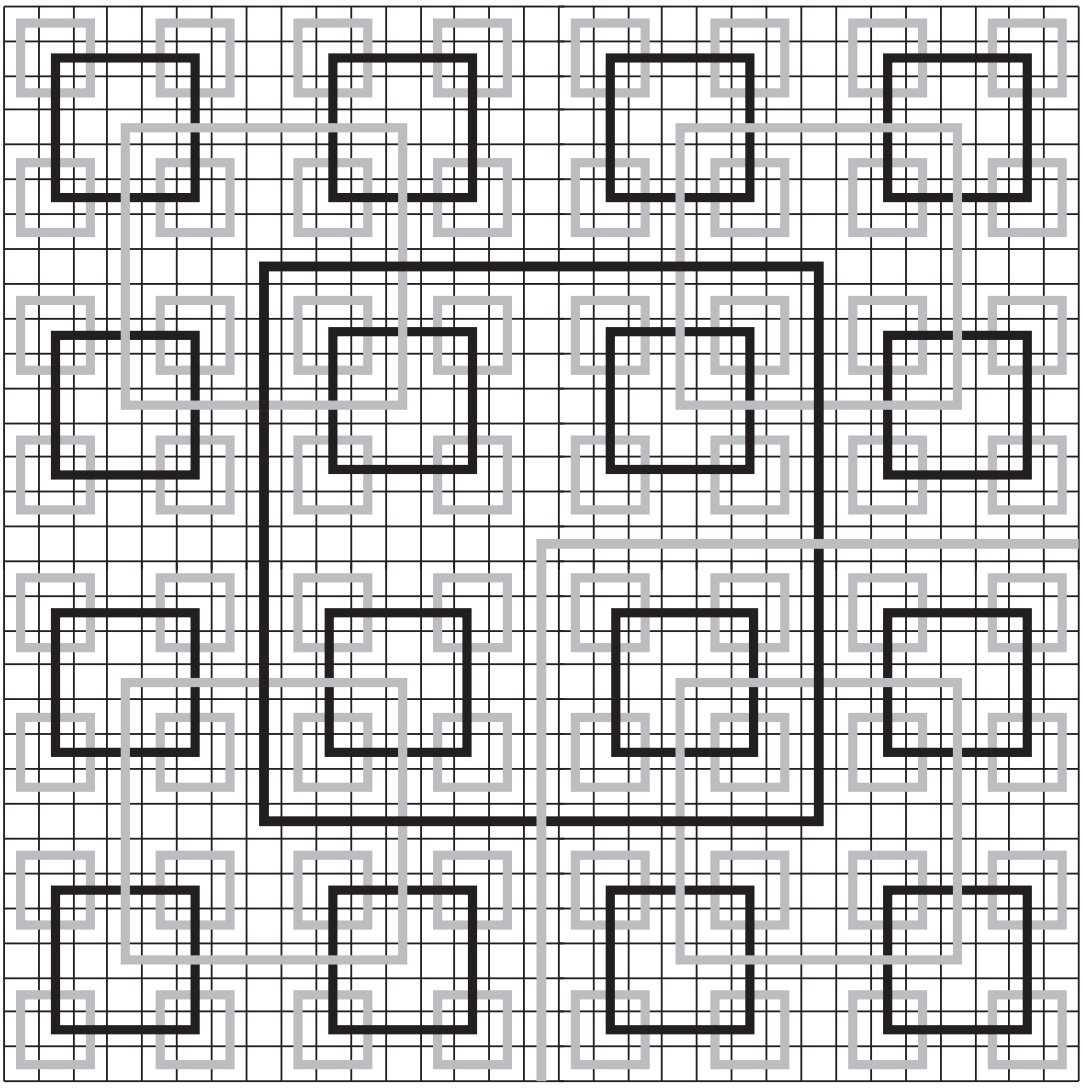
Notice that the side arrows form overlapping squares: The side arrows emitted from the crosses form squares whose centers contain crosses, which in turn are corners of bigger squares. The smallest squares have the corners at the odd-odd -positions. They are of size 2×2, and they only intersect one 4×4 square whose corner is at the center. Any other square S is of size $2^n \times 2^n$, for $n \geq 2$, and it intersects

one bigger square of size $2^{n+1} \times 2^{n+1}$ whose corner is at the center of S, and four smaller squares of sizes $2^{n-1} \times 2^{n-1}$ whose centers are the corners of of S.

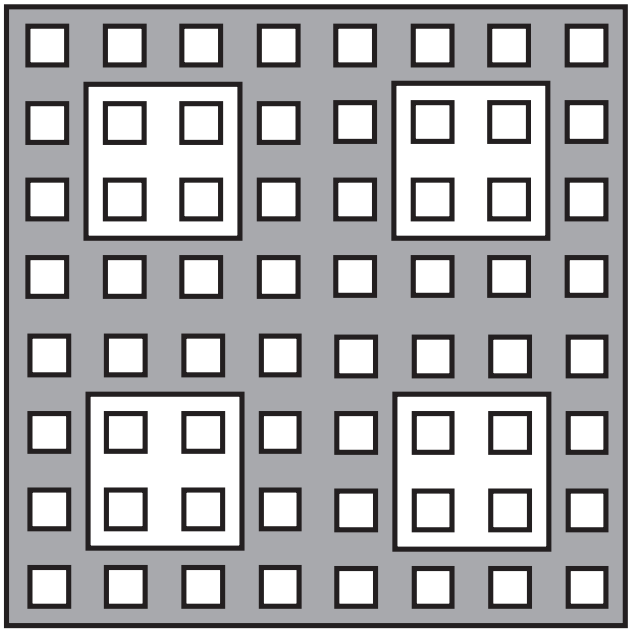
In order to pick non-overlapping squares we color the side arrows red or green according to the following rules:

- The side arrows of each cross are both red or both green. The crosses at odd-odd positions have green side arrows.
- In each arm the horizontal side arrows have the same color and the vertical side arrows have the same color. In this way the color is transmitted unchanged through the arm. If the arm contains both horizontal and vertical side arrows then these side arrows have different colors.
- In neighboring tiles the matching rule is that the meeting arrow heads and tails must have the same color.

Following these rules, each square will be colored completely red or green, and intersecting squares have opposite colors. The smallest squares are green, so the coloring of the squares is completely determined. Notice that red squares do not intersect each other and green squares do not intersect each other. The red squares are of sizes $4^n \times 4^n$ for $n = 1, 2, \ldots$. In the following figure green and red borders are indicated light and dark, respectively.



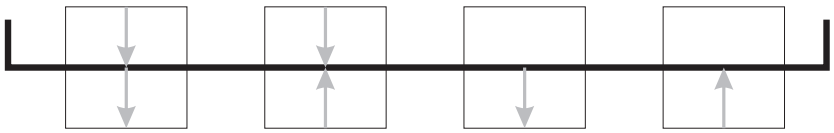
A piece of a Wang tiling with a seed tile will be enforced inside each red square. Small red squares nested within a larger red square contain their own copies. We call a region within a red border but outside all nested red borders within it a *board*. Here is the board with side $4^3 = 64$:



Next we need to identify those rows and columns of a board that run completely across the board without intersecting a smaller board inside. Let us call these *free* rows and columns. Let F_n be the number of free rows (and columns) in a board of side 4^n. In the middle the board of side 4^{n+1} we have the same smaller boards as in the middle of the 4^n board, and at the sides we have halves of those same boards. The boundary of the 4^{n+1} square occupies one row, so the total number of free rows in the 4^{n+1} board is $F_{n+1} = 2F_n - 1$. Since $F_1 = 3$, we easily obtain $F_n = 2^n + 1$. Hence a valid tiling necessarily contains boards with arbitrarily large numbers of free rows and columns.

To identify tiles of the board that are on a free row and/or a free column, we use a new set of arrows, called *obstruction signals*. There are vertical and horizontal obstruction signals, used to identify free columns and rows, respectively. An outer edge of a red border must emit or absorb an obstruction signal, whereas the inner edges of the red borders may absorb but not emit such a signal. Inner edges can also be without an obstruction signal.

Here are the four possible combinations of obstruction signals on the lower boundary of a red square:

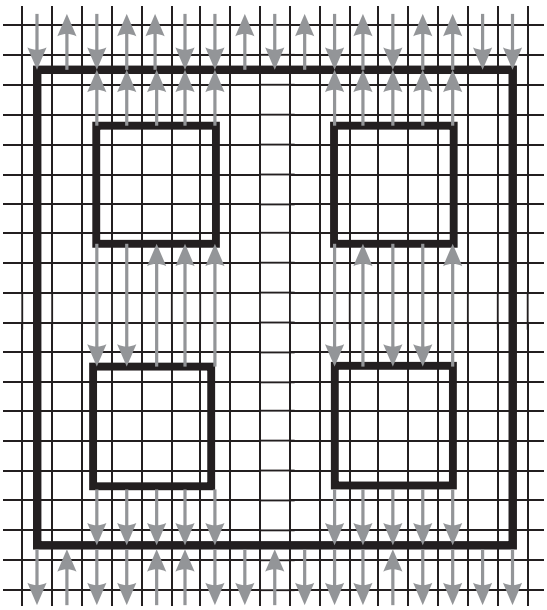


The other boundaries (top, left, right) are analogous. Note that the position of a red side arrow identifies whether it belongs to the top, bottom, left or right boundary of a red square.

Obstruction signals are transmitted unchanged through tiles that are not on the boundary of a board. Then no free column can contain a vertical obstruction signal and no free row can contain a horizontal obstruction signal, as such a signal would have to be emitted from the inside edge of the boundary. In contrast, any interior tile of a board that is not on a free column is either between the inner edge of the board and the outer edge of a smaller board, or between outer edges of two smaller boards. In either

case, there is a vertical obstruction signal at the tile. Analogously, any tile that is not on a free row must contain a horizontal obstruction signal. We conclude that the horizontal and vertical obstruction signals correctly identify the free rows and columns of the boards.

The following figure shows a possible arrangement of vertical obstruction signals on a 16×16 board:



Based on the presence or absence of horizontal/vertical obstruction signals, tiles inside a board can be classified into four classes:

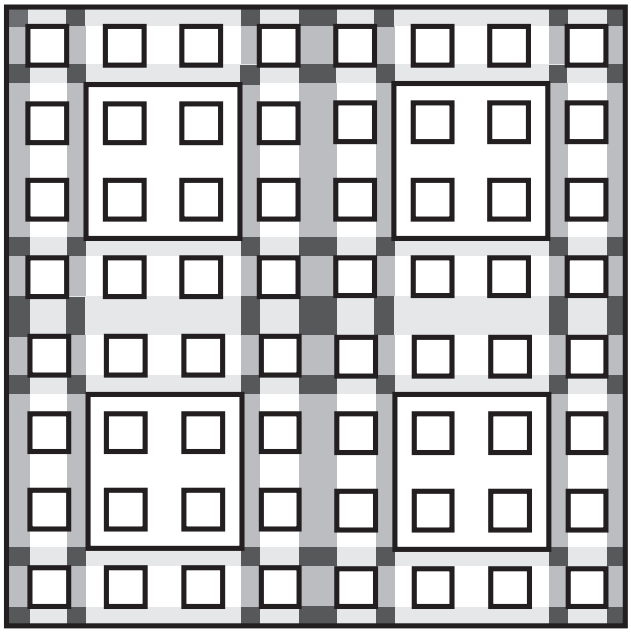
(00) with horizontal and with vertical obstruction signal,

(01) with horizontal and without vertical obstruction signal,

(10) without horizontal and with vertical obstruction signal,

(11) without any kind of obstruction signal.

Tiles of type (11) are free, and they form a scattered $(2^n + 1) \times (2^n + 1)$ -square, whose disjoint parts are connected by tiles of types (01) and (10):

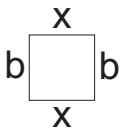


Let $\mathcal{R}$ be the tile set constructed above. Now we are ready to reduce the tiling problem with the seed tile into the tiling problem without a seed tile: Let $\mathcal{P}$ be a given set of Wang tiles, and let $s \in \mathcal{P}$ be the given seed tile. Let C be the set of colors used in $\mathcal{P}$. To determine if $\mathcal{P}$ admits a tiling that contains a copy of s we construct a set $\mathcal{X}$ of "sandwich tiles" (r, p), whose first component $r \in \mathcal{R}$, and the second component p is a Wang tile over the color set $C \cup \{b\}$, where $b \notin C$ is a new "blank color".

The first components tile according to the local matching constraints of $\mathcal{R}$ described above. The second components tile under the color constraints, as in Wang tiles. The set $\mathcal{X}$ contains all the pairs (r, p) that satisfy the following:

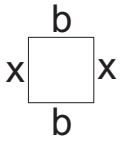
(a) If r is a free tile (no obstruction signal present in r) then $p \in \mathcal{P}$.

(b) If r is on a free column but not on a free row then p is a tile



where $x \in C$ and b is the blank color.

(c) If r is on a free row but not on a free column then p is a tile



where $x \in C$ and b is the blank color.

(d) If r is not on a free row or column then p is arbitrary: any element of $(C \cup \{b\})^4$ is acceptable.

(e) If r is a corner of a green square of size $> 2 \times 2$ (that is, a cross with green side arrows in an even-even position), then $p = s$, the seed tile.

Notice that the corners of green squares that are in even-even positions are exactly at the centers of red squares. These center positions are always free.

Condition (e) guarantees that the center of each board is paired with the seed tile s. Properties (b) and (c) mean that color information is transmitted along free rows and columns between disjoint parts of the board, while (a) guarantees that on free areas a tiling by $\mathcal{P}$ is formed. These conditions make the board behave as if the free rows and columns were contiguous and the board then is like a square board of side $2^n + 1$. Condition (d) simply allows different boards to be joined arbitrarily.

Let us prove that our sandwich tiles admit a tiling if and only if $\mathcal{P}$ admits a tiling that contains a copy of the seed tile s:

$\Longleftarrow$ Suppose first that $\mathcal{P}$ admits a tiling that contains s. Then we can properly tile any board by placing s at the center and scatter the proper tiling containing s in the free areas. Smaller nested boards can be tiled in the same way. Different boards are not immediate neighbors of each other since there are at least the red boundary tiles between them. So different boards can be tiled independently of each other. As we can tile arbitrarily large squares in this way, the whole plane can be tiles as well.

$\Longrightarrow$ For the converse direction, suppose that the sandwich tiles admit a tiling. The underlying Robinson's tiling necessarily contains special $(2^n - 1)$ -squares for arbitrarily large number n. Hence there are red squares of size $4^n \times 4^n$, for every n, and consequently there are arbitrarily large boards. The center of each board is paired with s, and the free areas of the board necessarily contain a piece of a valid tiling by $\mathcal{P}$. As the free area is arbitrarily large, and its center contains s, we conclude that $\mathcal{P}$ admits a tiling that contains a copy of tile s.

We have proved the following theorem by Berger. The proof presented here is from 1971 by Robinson.

Theorem 5.8 (Berger 1966) *The tiling problem "Does a given Wang protoset admit a valid tiling ?" is undecidable.* □

Each finite protoset is exactly one of the following types:

1. protosets that do not admit any tilings,
2. protosets that admit some periodic tilings,
3. aperiodic protoset

Membership in the first two classes are known to be semi-decidable (Lemmas 5.1 and 5.6). The undecidability of the tiling problem implies that membership in the third class cannot be semi-decidable, so there is no semi-algorithm to determine if a given protoset is aperiodic. Membership in the union of classes 2 and 3 is not semi-decidable (Theorem 5.8), and the membership in the union of classes 1 and 3 is not semi-decidable (Theorem 5.7). Hence we have been able to determine the semi-decidability status for all combinations of the three classes above.

5.6 The completion problem

Consider next the following decision problem: Is a given finite pattern part of some valid tiling ? We call it the completion problem. More precisely, a finite pattern over the tile set T is a pair (D, p) where $D \subseteq \mathbb{Z}^2$ is a finite domain, and $p : D \longrightarrow T$ assigns tiles to the cells in the domain. We say that (D, p) is a subpattern of configuration $c \in T^{\mathbb{Z}^2}$ if $c(\vec{n}) = p(\vec{n})$ for all $\vec{n} \in D$. We want to determine for a given (D, p) whether there exists a valid tiling that contains it as a subpattern.

The completion problem is clearly undecidable if the input contains both the tile set and the pattern (as the tiling problem with a seed tile is a particular case of this). But it turns out that the completion problem is undecidable already for some fixed tile sets. In this case only the pattern (D, p) is the input. A tile set with undecidable completion problem can be constructed from any Turing machine with

undecidable halting problem: here we consider computations of Turing machines from initial tapes that are not necessarily blank. Instead, we assume that there may be finitely many tape locations that have a non-blank symbol. So let us call *b-finite* a tape content $f : \mathbb{Z} \longrightarrow \Gamma$ such that the set $\{i \in \mathbb{Z} \mid f(i) \neq b\}$ is finite.

Theorem 5.9 *There exists a Turing machine $U = (S, \Gamma, \delta, s, h, b)$ such that the following decision problem is undecidable: "Given b-finite $f : \mathbb{Z} \longrightarrow \Gamma$, does U reach the halting state h from the initial configuration $(s, 0, f)$?"*

Proof. (sketch) The machine U we construct is a universal Turing machine: it is able to simulate any other Turing machine if a description of that machine is initially written on the tape. Then if we could determine whether U halts from a given initial tape then we could also determine for any given Turing machine whether it halts from the empty input tape.

First we observe that the halting problem from the empty tape is undecidable even among Turing machines with the binary tape alphabet $\Gamma = \{a, b\}$, where b is the blank symbol:

Lemma 5.10 *It is undecidable if a given Turing machine with binary tape alphabet eventually halts from the empty input tape.*

Proof. (sketch) For any given Turing machine M with k tape letters $1, 2, \ldots, k$, we effectively construct machine M' that has tape alphabet $\{0, 1\}$ and that halts from the empty input tape if and only if M halts from the empty input tape. On the tape of machine M' a binary encoding of the k letter alphabet is done using blocks of $n = \lceil \log_2(k) \rceil$ bits. The code for the blank tape letter of M is $00 \ldots 0$.

At all times, M' memorizes the current state q of machine M. To simulate one instruction of M, the machine M' does the following:

- It reads and memorizes the next block of n bits from the tape: this gives the current tape letter x of M.
- Let $\delta(q, x) = (p, y, d)$.
- Machine M' memorizes the new state p, writes the n bits representing y over the block that it just read, and moves n positions left or right on the tape, depending on the value of d.

This is the simulation loop for one step of M. The initial and halting states of M' are the first states of the simulation loop, with the initial and halting state of M being memorized, respectively.

It is clear that when started on the empty tape (0 is the blank), machine M' simulates M until, if ever, it halts. □

Based on the lemma, it is enough for our universal Turing machine U to simulate those Turing machines that have the tape alphabet $\{a, b\}$. The tape of U will contain five tracks:

1. On the first track, the tape of the simulated machine is written. This track has alphabet $\{a, b\}$.
2. The second track stores the state of the simulated machine. We may assume, without loss of generality, that the states of M are numbers $1, 2, \ldots, k$, where 1 is the initial state and k is the halting state. Current state i is expressed on track two as a segment of i symbols \$, starting at position zero of the tape. Outside the segment the track contains the blank b, so the track alphabet is $\{\$, b\}$.
3. The third track stores the position of M on its tape during the simulation. The track simply contains $\uparrow$ at cell i if machine M currently reads that cell. Other cells are blank. This track alphabet is $\{\uparrow, b\}$.

4. The fourth track stores the "program", i.e. the transition function δ of the simulated machine M. Each transition $\delta(i, x) = (j, y, d)$ is expressed as the word $\$^j\ y\ d$ where y is either a or b, and d is either L or R. For example, transition $\delta(i, x) = (4, a, L)$ is represented as $\$\$\$\$aL$. We represent the transition table δ as the word

$$\#\ E(1,a)\ \&\ E(1,b)\ \#\ E(2,a)\ \&\ E(2,b)\ \#\ \ldots\ \#\ E(k-1,a)\ \&\ E(k-1,b)\ \#$$

where $E(i, x)$ is the word representing $\delta(i, x)$. Note that the transitions from the halting state k are not written. The alphabet of track four is $\{\#, \&, \$, L, R, a, b, B\}$. (We use B for the blank.)

5. The fifth track is a "scratch pad" where temporary information is stored during the simulation.

As an example of the encoding of the program, consider the three-state Turing machine of Example 9, where we number the states as follows: $s = 1, t = 2, h = 3$. The corresponding program in our universal machine is

$$\#\ \$\$aL\ \&\ \$\$aR\ \#\ \$\$\$aL\ \&\ \$aL\ \#$$

The tape at the beginning of the simulation will be

			$																																
			↑																																
			#	$	$	a	L	&	$	$	a	R	#	$	$	$	a	L	&	$	a	L	#													

To simulate one step of M, the universal machine memorizes the current tape symbol scanned by M, as well as whether the position being scanned is positive or negative (so that ↑ can be easily found). Then it performs the following operations:

- Using the scratch pad (track 5) to store check markers, the machine matches the symbols $ of track 2 and the symbols # of track 4 starting from the left. This way the machine finds the i'th marker # from the program where i is the state of the simulated machine.
- If the symbol on track four that follows the i'th delimiter # is a blank then the machine halts. The simulated machine has namely entered the halting state k.
- If the tape letter being scanned by M is b then the machine finds the next & symbol on track 4. Now the machine is at the beginning of $E(i, x)$ for the next instruction to be executed.
- Using the scratch pad (track 5) the machine copies the signs $ from the word $E(i, x)$ on track 4 to track 2.
- The machine memorizes the the next two symbols y and d of the instruction. The machine finds the symbol ↑ on track 3, writes the symbol y in that location on track 1 and moves ↑ on track 3 one position left or right, depending on the value of d. The tape letter in the new location is memorized.
- The machine clears the scratch pad, and moves the position 0, ready to simulate the next step.

For example, after the first simulation round the tape in our sample machine will be

			a																																
			$	$																															
				↑																															
			#	$	$	a	L	&	$	$	a	R	#	$	$	$	a	L	&	$	a	L	#													

It is clear that each simulation loop of U sketched above properly simulates one step of M. If the simulated machine M enters the halting state, then also U halts.

If we had an algorithm to determine whether U halts when started on a given finite initial tape, then we could use this algorithm to determine if any given Turing machine M halts form the empty tape. Indeed, we construct the initial configuration where the program track four contains the description of M and ask whether U halts from this configuration. □

The universal Turing machine sketched in the proof above contains many states and and has a large tape alphabet. Smaller universal machines have been discovered.

Example 11. The following universal Turing machine (due to Y.Rogozhin) with $4 + 1$ states and 6 tape symbols has an undecidable halting problem, i.e. it satisfies the condition of the previous theorem: $M = (\{q_1, q_2, q_3, q_4, h\}, \{1, b, >, <, 0, c\}, \delta, q_1, h, b)$, and δ is given by the following table

	q_1	q_2	q_3	q_4
1	$(q_1, <, L)$	$(q_2, 0, R)$	$(q_3, 1, R)$	$(q_4, 0, R)$
b	$(q_1, >, R)$	$(q_3, >, L)$	$(q_4, <, R)$	(q_2, c, L)
$>$	(q_1, b, L)	$(q_2, <, R)$	(q_3, b, R)	$(q_4, <, R)$
$<$	$(q_1, 0, R)$	$(q_2, >, L)$	h	h
0	$(q_1, <, L)$	$(q_2, , 1, L)$	(q_1, c, R)	(q_2, c, L)
c	$(q_4, 0, R)$	(q_2, b, R)	$(q_1, 1, R)$	(q_4, b, R)

where the item on column q, row x is $\delta(q, x)$. □

Modifying slightly the Turing machine tile construction in Section 5.2 we can now easily obtain the following:

Theorem 5.11 *There exists a finite set $\mathcal{P}$ of Wang prototiles such that the following problem is undecidable: "Is a given finite pattern a subpattern of some valid tiling ?"*

Proof. In the homework assignments.

□

5.7 Beyond aperiodicity: arecursive tile sets

Robinson's tile set is aperiodic: no periodic tilings exist. Still, valid tilings that are "simple" exist. Using the special $2^n - 1$ -squares one can effectively construct bigger and bigger portions of a fixed valid tiling.

But there exist tile sets that only admit very complicated tilings: namely tilings that cannot be algorithmically constructed. We call a tiling $c : \mathbb{Z}^2 \longrightarrow T$ *recursive* if there exists an algorithm that outputs $c(i, j)$, when given arbitrary integers i, j as input. If no such algorithm exists then c is called *non-recursive*. Analogously, a tape content $f : \mathbb{Z} \longrightarrow \Gamma$ of a Turing machine is recursive if there exists an algorithm that returns $f(i)$ for any given input $i \in \mathbb{Z}$. Otherwise f is non-recursive.

Clearly any two-way periodic tiling is recursive. Some non-periodic tilings are recursive, too: for example Robinson's tile set admits a recursive, non-periodic tiling. Analogously to aperiodicity, we define the concept of *arecursivity* as follows: Wang tile set $\mathcal{P}$ is arecursive if and only if

(i) it admits valid tilings, and

(ii) it does not admit any recursive valid tilings.

Clearly any arecursive tile set is aperiodic, but the converse is not true since Robinson's tile set is not arecursive. So arecursivity is a stronger property than aperiodicity.

Arecursive tile sets exist. First, let us sketch a proof that there exist Turing machines $M_R = (S, \Gamma, \delta, s_0, s_h, b)$ with the following behavior:

(i) For every recursive $f : \mathbb{Z} \longrightarrow \Gamma$, machine M_R halts from the initial configuration $(s_0, 0, f)$, but

(ii) there exists a (non-recursive) tape content $f : \mathbb{Z} \longrightarrow \Gamma$ such that M_R does not halt from the initial configuration $(s_0, 0, f)$.

To construct such M_R we consider the problem of determining which of two halting states does a given Turing machine halt. In this problem we are given a TM M with two halting states h_1 and h_2. We are looking for an algorithm that returns answer 1 or 2 if M halts in state h_1 or h_2, respectively, when started on the blank tape. If M does not halt then the algorithm can (and must) return either answer 1 or answer 2. That such an algorithm can not exist can be proved using a diagonal argument (which we skip), similar to Turing's proof for the undecidability of the halting problem:

Lemma 5.12 *There is no algorithm that, for any given TM M with two halting states $h_1, h_2 \in S$,*

- *always returns an answer "1" or "2", and*
- *returns answer "1" ("2",respectively) if M halts in state h_1 (or h_2, respectively) when started on the blank tape.*

We say that the set A of Turing machines that halt in state h_1 is *recursively inseparable* from the set B of Turing machines that halt in state h_2. (More generally, disjoint sets A and B are recursively inseparable if there is no algorithm that returns an answer on every input x, and if $x \in A$ then the answer must be 1 and if $x \in B$ then the answer must be 2. In other terms: there is no decidable set R such that $A \subseteq R$ and $B \cap R = \emptyset$.)

Now we can construct TM M_R with several tracks on the tape:

- On the first track we have symbols of the alphabet $\{1, 2\}$. These symbols are not modified by M_R.
- On the second track the machine enumerates integers $n = 1, 2, 3, \ldots$ one-by-one.
- For each value n on the second track, the machine then enumerates on the third track numbers $m = 1, 2, \ldots n$. Values of m are interpreted as TM descriptions using, say, the encoding given in the proof of Theorem 5.9: Number m is written in the number system with 7 digits, using the symbols of the alphabet $\Sigma = \{\#, \&, \$, L, R, a, b\}$ as the digits.
- If m is not a proper encoding of a TM transition rule (which can be easily checked) then nothing is done, but the machine simply moves on to the next value of m.
- But if m represents some TM M then a simulation of M is started from the blank tape, using the universal machine from the proof of Theorem 5.9. The simulation is done only for n steps, where n is the number on the second track.
- If the simulated machine reaches state h_1 or h_2 (which we can fix to be the second and the third state of the machine, respectively) before n simulation steps are executed then M_R checks the symbol on the first track in position m: if the symbol is 1 but M reached state h_2, or the symbol is 2 but M reached state h_1, then M_R halts. Otherwise it moves on to the next value of m.

The idea of M_R is to simulate all Turing machines for arbitrarily long times, and to verify for each machine that the first track correctly identifies which state h_1 or h_2 is first reached in each machine. Machine M_R halts from an initial configuration if and only if for some Turing machine m the first track identifies in position m incorrectly the state in which m halts.

Suppose the initial tape content $f : \mathbb{Z} \longrightarrow \Gamma$ of machine M_R is recursive. Let us prove that M_R must halt. Suppose the contrary: M_R does not halt. Then for any Turing machine m we can effectively calculate the symbol on the first track in position m of f. Because M_R does not halt, this symbol must be 1 if m halts in state h_1 and 2 if m halts in state h_2. So we have described an algorithm that contradicts lemma 5.12. We conclude that M_R halts when started on any recursive initial tape content.

On the other hand, if the first track is such that it correctly identifies which machines halt in state h_1 and h_2 then M_R does not halt. So there are (non-recursive) initial tapes from which M_R does not halt.

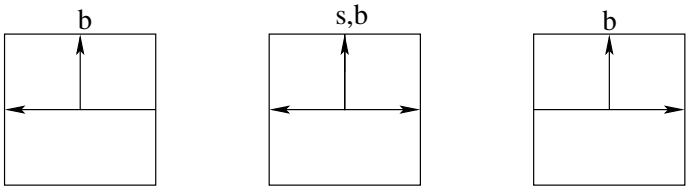
We have sketched a proof of the following lemma:

Lemma 5.13 *There exists a Turing machine M_R that halts when started on any recursive initial tape, but for some non-recursive initial tape M_R does not halt.* □

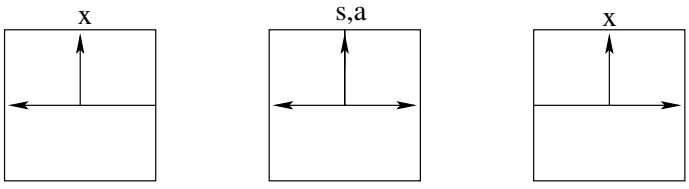
Based on machine M_R we can now easily construct a Wang tile set with a seed tile such that there are valid tilings that contain the seed tile, but none of these tilings is recursive. This is a weaker property than arecursivity defined above, because of the presence of a seed tile.

Theorem 5.14 *There exists a finite set $\mathcal{P}$ of Wang prototiles such that for some $t \in \mathcal{P}$ every valid tiling that contains t is non-recursive, and there are valid non-recursive tilings that contain t.*

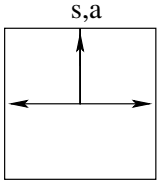
Proof. Let us construct for M_R of lemma 5.13 the machine tiles of Section 5.2, except that the start tiles



that represent the initial empty tape will be replaced by the following tiles for all tape letters $x \in \Gamma$, the initial state $s \in S$, and a single tape letter $a \in \Gamma$. Letter a is chosen so that there is an initial tape content $f : \mathbb{Z} \longrightarrow \Gamma$ with $f(0) = a$ such that M_R does not halt when started on tape f.



These start tiles allow non-blank initial tapes. As the seed tile t we select the new start tile



The initialization tiles allow the horizontal row with t to contain any initial tape content f with $f(0) = a$. The machine tiles then force the rows above to simulate machine M_R. If M_R halts then the tiling becomes impossible.

These tiles admit a valid tiling: We simply choose an initial row that represents a tape content f from which M_R does not halt. But the tiles do not admit any recursive tiling containing t: If the tiling is recursive then the horizontal row containing t is recursive, so we would have a recursive initial tape from which M_R does not halt, a contradiction with Lemma 5.13. □

It is possible (but we skip the proof) to modify Robinson's tile set so that the seed tile restriction in Theorem 5.14 can be removed. As in the proof of Theorem 5.8 we form nested boards, and a simulation of the Turing machine M_R is done on all boards. The main problem to be addressed in the construction is the fact the simulations on all boards should be from the same initial tape content. Otherwise, if different boards are allowed to run M_R on different initial tape contents then a recursive tiling could be easily built. Hence new signals need to be introduced that carry the information about the initial tape content between boards of different sizes, so that different boards are forced to be consistent with each other.

Theorem 5.15 *There exist arecursive sets of Wang tiles.*

Proof. For the original proof, if interested, see:

Dale Myers. Nonrecursive Tilings of the Plane, II. *The Journal of Symbolic Logic*, **39(2)**, pp. 286-294, 1974. □

6 Compact topology on Wang tilings

In this section we assign a metric to the space $T^{\mathbb{Z}^2}$ of configurations over the finite tile set T. The space under this metric is compact and complete. Convergence of a sequence $c_1, c_2, \ldots$ of elements under this metric is exactly equivalent to the convergence introduced in Section 4.2. The compactness principle of that section then simply reflect the compactness of the metric space.

We define the distance $d(e, c)$ between configurations $e, c \in T^{\mathbb{Z}^2}$ as follows:

$$d(e,c) = \begin{cases} 0, & \text{if } e = c, \\ 2^{-\min\{|x|+|y| \;\mid\; e(x,y) \neq c(x,y)\}}, & \text{if } e \neq c. \end{cases}$$

In other words, two configurations that differ in a cell that is close to $\vec{0}$ are far away from each other under this metric, while configurations that agree with each other on a large area around the origin are close to each other. Under this metric, two configurations have distance $< 2^{-r}$ if and only if they agree with each other at all positions (x, y) where $|x| + |y| \leq r$.

Note that other vector norms $\|(x, y)\|$ could be used instead $|x|+|y|$, and any other decreasing function could be used instead of $x \mapsto 2^{-x}$. A different metric, but the same topology would result.

Lemma 6.1 *Function $d : T^{\mathbb{Z}^2} \times T^{\mathbb{Z}^2} \longrightarrow \mathbb{R}$ is a metric.*

Proof. We have to check the three defining properties of metric:

(a) $d(c, e) \geq 0$, and $d(c, e) = 0$ if and only if $c = e$,

(b) $d(c, e) = d(e, c)$, and

(c) $d(c,e) \leq d(c,c') + d(c',e)$.

The first two conditions (a) and (b) are immediate. The third condition (c), called the triangle inequality, follows from the fact that for every $\vec{x} \in \mathbb{Z}^2$, if $c(\vec{x}) \neq e(\vec{x})$ then either $c(\vec{x}) \neq c'(\vec{x})$ or $c'(\vec{x}) \neq e(\vec{x})$, or both. This means that either $d(c,c') \geq d(c,e)$ or $d(c',e) \geq d(c,e)$, so even the strong form

$$d(c,e) \leq \max\{d(c,c'), d(c',e)\}$$

of the triangle inequality holds. □

From now on we consider $T^{\mathbb{Z}^2}$ as a metric topological space under this metric. The following subsection contains a brief review of some basic facts about metric spaces.

6.1 Review of topology and metric spaces

Let X be a set. A family $\mathcal{T}$ of subsets of X is called a *topology* if it satisfies the following three conditions:

(i) $\emptyset \in \mathcal{T}$ and $X \in \mathcal{T}$,

(ii) the union of the sets in any subfamily of $\mathcal{T}$ is in $\mathcal{T}$,

(iii) the intersection of finitely many elements of $\mathcal{T}$ is always in $\mathcal{T}$.

Elements of $\mathcal{T}$ are called *open* sets, and their complements (with respect to X) are *closed* sets. A set that is both open and closed is called *clopen*.

Example 12. For any X, let $\mathcal{T}$ contain all subsets of X. Then $\mathcal{T}$ is a topology, the *discrete* topology of X. Also $\{X, \emptyset\}$ is a topology, the *trivial* topology of X. □

Example 13. Let us call $S \subseteq \mathbb{R}$ open if for every $x \in S$ there is a positive real $\varepsilon > 0$ such that $|y - x| < \varepsilon \Longrightarrow y \in S$. These open sets form a topology of $X = \mathbb{R}$. It is called the usual topology of $\mathbb{R}$. For example, all open intervals (a,b) for $a < b$ are open sets. Closed intervals $[a,b]$ are not open but they are closed. Set $\mathbb{Q}$ of rational numbers is not open or closed. The only clopen sets are $\emptyset$ and $\mathbb{R}$. □

Generalizing the previous example, let X be a set and let $d : X \times X \longrightarrow \mathbb{R}$ be a metric. For every $\varepsilon > 0$ and $x \in X$ we denote

$$B_\varepsilon(x) = \{y \in X \mid d(x,y) < \varepsilon\}$$

and call $B_\varepsilon(x)$ the (open) ε-ball with center x. Let us call $U \subseteq X$ open if

$$\forall x \in U \ : \ \exists \varepsilon > 0 \ : \ B_\varepsilon(x) \subseteq U.$$

These open sets form a topology of X, the *metric topology* induced by d.

Example 14. The discrete topology is induced by the discrete metric

$$d(x,y) = \begin{cases} 0, & \text{if } x = y, \\ 1, & \text{if } x \neq y. \end{cases}$$

In contrast, if $|X| \geq 2$ then the trivial topology $\{X, \emptyset\}$ is not metric. □

Let $A \subseteq X$. Point $x \in X$ is an *accumulation point* of A if every open set U that contains x also contains some element $y \neq x$ of A. The following simple properties hold for closed sets:

Proposition 6.2 *A subset $A \subseteq X$ is closed if and only if its accumulation points belong to A. Closed sets satisfy the following properties (that are dual statements of the defining properties of open sets):*

(i) The empty set $\emptyset$ is closed, and X is closed,

(ii) the intersection of any number of closed sets is closed, and

(iii) the union of a finite number of closed sets is closed.

□

Let $A \subseteq X$. The *closure* of A is the intersection of all closed sets that contain A. It is then the smallest closed set that contains A. We denote the closure of A by $\overline{A}$. Notice that A itself is closed if and only if $\overline{A} = A$. Notice also that the closure of A is the union of A and its set of accumulation points.

Set A is called *dense* if $\overline{A} = X$.

Example 15. Consider the usual topology of $\mathbb{R}$. All real numbers are accumulation points of the set $\mathbb{Q}$ of rational numbers. This means that the closure of $\mathbb{Q}$ is $\mathbb{R}$, so $\mathbb{Q}$ is dense in $\mathbb{R}$. Accumulation points of the open interval $(0, 1)$ are the elements of the closed interval $[0, 1]$, while the set $\mathbb{Z}$ of integers has no accumulation points. □

Let $A \subseteq X$. Point $x \in A$ is an *interior point* of A if there is an open set U such that $x \in U$ and $U \subseteq A$. The set of all interior points of A is the *interior* of A. It is easily seen to be the union of all open subsets of A, or equivalently, the largest open subset of A. Then set A is open if and only if its interior is A itself.

The *exterior* of set $A \subseteq X$ is the interior of the complement of A, and the *boundary* of A consists of all points that are not in the interior or the exterior of A. Note that the interior, exterior and boundary of A is a partitioning of X. A set $A \subseteq X$ is called a *neighborhood* of $x \in X$ if x is an interior point of A, that is, if there is an open set U such that $x \in U \subseteq A$.

Example 16. In the usual topology of $\mathbb{R}$, the interior, exterior and the boundary of an open interval (a, b) are (a, b), $(-\infty, a) \cup (b, \infty)$ and $\{a, b\}$, respectively. The closed interval $[a, b]$ has these same interior, exterior and boundary. Set $\mathbb{Q}$ has empty interior and exterior. All real numbers are in its boundary. □

A topology is called *Hausdorff* if for every $x \neq y$ there are open U_x and U_y such that $x \in U_x$, $y \in U_y$ and $U_x \cap U_y = \emptyset$. In other words, any two distinct points have non-intersecting neighborhoods.

Example 17. Every metric topology is Hausdorff. Indeed, if $x \neq y$ then $d(x, y) > 0$. If we choose $\varepsilon = \frac{1}{2} d(x, y)$ then $B_\varepsilon(x)$ and $B_\varepsilon(y)$ are non-intersecting neighborhoods of x and y □

A sequence $x_1, x_2, \ldots$ of points of X *converges* to point $x \in X$ if for every open $U \subseteq X$ that contains x there is positive integer n such that $x_i \in U$ for all $i \geq n$. If the topology is metric this is equivalent to saying that for every $\varepsilon > 0$ there is n such that $d(x_i, x) < \varepsilon$ for all $i \geq n$.

Note that generally a converging sequence may converge to several different points, but if the topology is Hausdorff (e.g. metric) the limit is unique.

Proposition 6.3 *In Hausdorff topology every converging sequence converges to a unique point.* □

Proof. Suppose $x_1, x_2, \ldots$ converges to x and y where $x \neq y$. Since X is Hausdorff, there are open sets U and V such that $x \in U$, $y \in V$ and $U \cap V = \emptyset$. By the definition of convergence, $x_i \in U$ and $x_i \in V$ for all sufficiently large i, a contradiction. □

Note: the proposition does not hold in all topological spaces. For example, in the trivial topology $\mathcal{T} = \{\emptyset, X\}$ every sequence converges to every point.

In Hausdorff topology we denote by $\lim_{i \to \infty} x_i$ the unique point into which the sequence $x_1, x_2, \ldots$ converges, if it exists. This point is the limit of the sequence.

The following proposition states that if the topology is metric then the closure $\overline{A}$ of any set A consists exactly of the limits of converging sequences of elements of A:

Proposition 6.4 *Let X be a metric space and $A \subseteq X$. Then $x \in \overline{A}$ if and only if $x = \lim_{i\to\infty} a_i$ for some converging sequence $a_1, a_2, \ldots$ where all $a_i \in A$.*

Proof. "$\Longleftarrow$": Let $a_1, a_2, \ldots$ be a converging sequence where all $a_i \in A$ and let $x = \lim_{i\to\infty} a_i$. Let U be an arbitrary open set that contains x. By the definition of convergence there are some $a_i \in U$, so $U \cap A \neq \emptyset$. This means that $x \in \overline{A}$. (This direction of the proof holds for any topological space.)

"$\Longrightarrow$": Conversely, suppose $x \in \overline{A}$. For every positive integer i, let a_i be an element of $A \cap B_{\frac{1}{i}}(x)$. Then $d(x, a_i) < \frac{1}{i}$, so $x = \lim_{i\longrightarrow\infty} a_i$. □

Corollary 6.5 *In metric space X, set A is closed if and only if it contains the limit of every converging sequence of its elements.* □

A family $\mathcal{B}$ of open sets is called a *base* of the topology iff every open set is the union of some members of $\mathcal{B}$. Equivalently: $\mathcal{B} \subseteq \mathcal{T}$ is a base if for every open set U and $x \in U$ there exists some $B \in \mathcal{B}$ with the property that $x \in B \subseteq U$.

Example 18. The open intervals (a, b) with $a < b$ form a base of the usual topology of $\mathbb{R}$. More generally, in any metric topology the open balls $B_\varepsilon(x)$ over all $\varepsilon > 0$ and $x \in X$ form a base. □

If $\mathcal{B}$ is a base of a topology then this topology is uniquely determined by $\mathcal{B}$:open sets are exactly the unions of members of $\mathcal{B}$.

Next we define compactness. Let $A \subseteq X$ where X is a topological space. A family of open sets U_i is called an *open cover* of A if every element of A belongs to some U_i. A subfamily of an open cover of A is called a *subcover* if it is also a cover of A.

Set $A \subseteq X$ is called *compact* if every open cover of A has a finite subcover of A. The topology is called compact if the whole space X is compact. In other words, a topology is compact if every family of open sets whose union is X has a finite subfamily whose union is X.

Example 19. In the usual topology of $\mathbb{R}$ the set

$$A = \{0\} \cup \{\frac{1}{n} \mid n \in \mathbb{Z}_+\}$$

is compact. Namely, an open set that contains 0 covers all but finitely many elements of A. So any open cover of A contains a finite subcover: Open set U that covers 0 together with a finite number of open sets that cover the finitely many elements of A that are outside of U.

On the other hand, set $B = \{\frac{1}{n} \mid n \in \mathbb{Z}_+\}$ is not compact. It has an open cover in which every open set covers exactly one element of B. Such cover has no finite subcover. □

The following proposition states the finite intersection property. It is dual to the open cover property we used as the definition, and in fact the finite intersection property could have been taken equally well as the definition of compactness. We state the property for the whole space X:

Proposition 6.6 *Topology of X is compact if and only if every family of closed sets whose intersection is empty has a finite subfamily whose intersection is empty.*

Proof. This follows directly from the definition of compactness and de Morgan's laws: A family of open sets is a cover of X if and only if the family of their complements have empty intersection. □

We typically apply the previous proposition in the following set-up:

Corollary 6.7 *Let $F_1 \supseteq F_2 \supseteq F_3 \supseteq \dots$ be an infinite chain of closed sets in a compact space X. If*

$$\bigcap_{i=1}^{\infty} F_i = \emptyset,$$

then $F_i = \emptyset$ for some i. □

The next proposition gives a characterization of compact subsets in metric spaces. The proposition gives a condition that looks very similar to Proposition 4.2 for configurations. In fact, we use the proposition later to show the compactness of the configuration space. The proposition is valid (and is stated) for arbitrary metric spaces, but we only prove it now for metric spaces that have a countable base. Our configuration space satisfies this restriction, so the proof is sufficient for our set-up. The proof for general metric spaces is not very difficult either.

Proposition 6.8 *Suppose X is a metric space. Set $A \subseteq X$ is compact if and only if every sequence $a_1, a_2, \dots$ of elements of A has a subsequence that converges to an element of A.*

Proof. "$\Longrightarrow$" Suppose A is compact, and let $a_1, a_2, \dots$ be arbitrary sequence where each $a_i \in A$.

Suppose first that there is some $a \in A$ such that for every $\varepsilon > 0$ the ball $B_\varepsilon(a)$ contains infinitely many different elements of the sequence $a_1, a_2, \dots$. Then the sequence has a subsequence that converges to a: There namely is a subsequence whose n'th element belongs to $B_{\frac{1}{n}}(a)$.

Suppose then that for every $a \in A$ there is some $\varepsilon_a > 0$ such that $B_{\varepsilon_a}(a)$ only contains finitely many different elements of the sequence $a_1, a_2, \dots$. Clearly the family of $B_{\varepsilon_a}(a)$ over all $a \in A$ is an open cover of A, so by compactness of A it has a finite subcover

$$U_i = B_{\varepsilon_{a_i}}(a_i) \text{ for } i = 1, 2, \dots m.$$

But each U_i only covers finitely many different elements of sequence $a_1, a_2, \dots$, while each element of the sequence is covered by some U_i. This means that the sequence has only finitely many different elements. Then some element $a \in A$ repeats infinitely many times in the sequence so the sequence has a constant subsequence $a, a, \dots$ which trivially converges to $a \in A$.

"$\Longleftarrow$" Suppose every sequence of elements of A has a converging subsequence whose limit is in A. Here we simplify the set-up by making the additional assumption that the topology has a countable base. Then it is enough to show that any countable open cover of A has a finite sub-cover. (Indeed, for an arbitrary open cover by U_i we can consider instead the countable cover that consists of all base sets B_j that are completely included in some U_i. If every countable cover has a finite subcover, then the original cover also has a finite subcover where we take for each selected B_j one U_i from the original cover that satisfies $B_j \subseteq U_i$.)

So consider a countable open cover $\{U_1, U_2, \dots\}$ of A. If it has no finite subcover then for every i there is some $a_i \in A$ such that $a_i \notin U_j$ for all $j < i$. By the hypothesis, sequence $a_1, a_2, \dots$ has a converging subsequence with limit $a \in A$. But $a \in U_j$ for some j, and then by the definition of convergence $a_i \in U_j$ for infinitely many indices i. In particular, there is $i > j$ such that $a_i \in U_j$, which contradicts the choice of a_i's. We conclude that a finite subcover must exist. □

Next two propositions show that in our forthcoming situation compact sets of the space are exactly the closed sets.

Proposition 6.9 *If X is a compact topological space then every closed $A \subseteq X$ is compact.*

Proof. Let $A \subseteq X$ be closed. Consider an open cover of A. Together with the complement of A it forms an open cover of X. By compactness of X this has a finite subcover of X, from which we obtain a finite subcover of A by removing the complement of A (if present). Hence A is compact. □

Proposition 6.10 *If X is Hausdorff then every compact $A \subseteq X$ is closed.*

Proof. Let $A \subseteq X$ be compact. Let $x \in X \setminus A$. By the Hausdorff property, for every $a \in A$ there are open sets U_a and V_a such that $a \in U_a$, $x \in V_a$ and $U_a \cap V_a = \emptyset$. Sets U_a form an open cover of A so by compactness of A there is a finite subcover $U_{a_1}, \ldots, U_{a_m}$ of A. But then the intersection

$$V_x = V_{a_1} \cap \cdots \cap V_{a_m}$$

of the corresponding sets V_{a_i} is an open set satisfying $x \in V_x$ and $V_x \cap A = \emptyset$. The union of sets V_x over all $x \in X \setminus A$ is the complement of A. Since the union is open, we see that A is closed. □

A topological space is *separable* if it has a countable dense subset,and it is *second countable* if it has a countable base. Our space of interest is both separable and second countable. In fact, every compact metric space has these properties.

Proposition 6.11 *A compact metric space is separable.*

Proof. For every n the cover of X by the open balls $B_{1/n}(x)$ has a finite subcover. The centers of all the balls in these finite subcovers for $n = 1, 2, 3, \ldots$ form a countable set A. It is dense: For every $y \in X$ and $n \geq 1$ there is a ball $B_{1/n}(x)$ with center $x \in A$ that contains y. Then $x \in B_{1/n}(y)$. □

Proposition 6.12 *A metric space is separable if and only if it has a countable base.*

Proof. Let $\{x_1, x_2, \ldots\}$ be a dense countable subset of X. Then the open balls $B_{1/n}(x_i)$ over all positive integers i, n form a countable base. Indeed: For every open U and $x \in U$ there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq U$. Choose an integer $n > 2/\varepsilon$. Some $x_i \in B_{1/n}(x)$. Because $1/n < \varepsilon/2$ we have

$$x \in B_{1/n}(x_i) \subseteq B_\varepsilon(x) \subseteq U.$$

Conversely, if $U_1, U_2, \ldots$ is a countable base, then $\{x_1, x_2, \ldots\}$ is dense where each $x_i \in U_i$. □

Let $A \subseteq X$. We say that point $x \in A$ is *isolated* in A if there is an open set U such that $A \cap U = \{x\}$. In other words, some open neighborhood of x does not contain any other elements of A. A non-empty set S is called *perfect* if it is closed and has no isolated points.

Proposition 6.13 *In a compact metric space, a perfect set is uncountable.*

Proof. Clearly, in a Hausdorff space, all points of a finite set are isolated, so a perfect set is infinite. Suppose there is a countable perfect set

$$S = \{x_1, x_2, \ldots\}.$$

In the following we define a decreasing sequence $F_1 \supseteq F_2 \supseteq F_3 \supseteq \ldots$ of closed sets such that, for all i, set F_i contains an open neighborhood U_i of some element of S, but $x_i \notin F_i$.

First, $F_1 = \overline{B}_\varepsilon(x_2)$ where $\varepsilon < d(x_1, x_2)$. Then the open neighborhood $U_1 = B_\varepsilon(x_2)$ of x_2 is contained if F_1, but $x_1 \notin F_1$. Suppose then F_{i-1} and U_{i-1} have been defined. Because S has no isolated points, the open neighborhood of any point contains also other elements of S. Hence U_{i-1} contains at least two elements of S, and consequently some $a \in S$, $a \neq x_i$, is in U_{i-1}. We choose

$$\begin{aligned} F_i &= F_{i-1} \cap \overline{B}_\varepsilon(a) \text{ and} \\ U_i &= U_{i-1} \cap B_\varepsilon(a), \end{aligned}$$

where $\varepsilon < d(a, x_i)$. Then F_i is closed, $F_{i-1} \supseteq F_i$ and $x_i \notin F_i$. Moreover, U_i is open and $a \in U_i \subseteq F_i$.

Because $F_i \cap S$ are closed and non-empty, by Corollary 6.7 the intersection

$$A = \bigcap_{i=1}^{\infty} F_i \cap S$$

is not empty. But $x_i \notin F_i$, so $A = \emptyset$, a contradiction. □

Finally, a few words about continuous functions. Let X and Y be two topological spaces. A function $f : X \longrightarrow Y$ is *continuous* at point $x \in X$ if for every open $V \subseteq Y$ that contains $f(x)$ there exists an open $U \subseteq X$ such that $x \in U$ and $f(U) \subseteq V$.

If X and Y are metric spaces with metrics d and e, respectively, then continuity at x is equivalent to the following: For every $\varepsilon > 0$ there exists $\delta > 0$ such that $f(B_\delta(x)) \subseteq B_\varepsilon(f(x))$.

We call function $f : X \longrightarrow Y$ is *continuous* if it is continuous at every $x \in X$.

Example 20. If X has the discrete topology then every function $f : X \longrightarrow Y$ is continuous. Also, if Y has the trivial topology $\{\emptyset, Y\}$ then every $f : X \longrightarrow Y$ is continuous. In all topological spaces X and Y all constant functions $f : X \longrightarrow Y$ are continuous. If X has the trivial topology and Y has the discrete topology then the constant functions are the only continuous functions. □

Proposition 6.14 *Let $f : X \longrightarrow Y$ be a function between two topological spaces. The following conditions are equivalent:*

(i) Function $f : X \longrightarrow Y$ is continuous,

(ii) pre-image $f^{-1}(V)$ is open in X for every open $V \subseteq Y$,

(iii) pre-image $f^{-1}(C)$ is closed in X for every closed $C \subseteq Y$.

Proof. (i) $\Longrightarrow$ (ii): Suppose f is continuous and let $V \subseteq Y$ be open. Let $x \in f^{-1}(V)$ be arbitrary, so $f(x) \in V$. From continuity it follows that there is an open $U \subseteq X$ such that $f(U) \subseteq V$ and $x \in U$. This means that $x \in U \subseteq f^{-1}(V)$, which implies that $f^{-1}(V)$ is open.

(ii) $\Longrightarrow$ (i): Suppose $f^{-1}(V)$ is open for every open $V \subseteq Y$. Let $x \in X$ be arbitrary. Let us show that f is continuous at point x. Let $f(x) \in V$ for open $V \subseteq Y$. Then $U = f^{-1}(V)$ is an open set that satisfies $x \in U$ and $f(U) \subseteq V$. So f is continuous at x.

(ii) $\Longleftrightarrow$ (iii): Follows directly from the fact that for every $A \subseteq Y$ holds

$$X \setminus f^{-1}(A) = f^{-1}(Y \setminus A).$$

□

Next propositions give some properties of continuous functions and compact sets.

Proposition 6.15 *Suppose function $f : X \longrightarrow Y$ is continuous. For every compact A the set $f(A)$ is compact.*

Proof. Consider an open cover of $f(A)$ by open sets V_i. Then, by Proposition 6.14 the sets $f^{-1}(V_i)$ form an open cover of A. By compactness of A there is a finite subcover of A by $f^{-1}(V_i)$ where $i \in F$ for some finite set F. But then the corresponding sets V_i for $i \in F$ form a finite subcover of $f(A)$. Hence $f(A)$ is compact. □

Proposition 6.16 *If $f : X \longrightarrow Y$ is a continuous bijection where X is compact and Y is Hausdorff then the inverse function $f^{-1} : Y \longrightarrow X$ is also continuous.*

Proof. By Proposition 6.14 it is enough to show that for every closed $A \subseteq X$ also $f(A)$ is closed. But if $A \subseteq X$ is closed then by Proposition 6.9 it is also compact. By Proposition 6.15 set $f(A)$ is also compact, and then by Proposition 6.10 set $f(A)$ is closed. □

6.2 Basic facts about the configuration space

Let us return to the space of interest to us: The space $T^{\mathbb{Z}^2}$ of configurations, with the metric

$$d(e,c) = \begin{cases} 0, & \text{if } e = c, \\ 2^{-\min\{|x|+|y| \;\mid\; e(x,y)\neq c(x,y)\}}, & \text{if } e \neq c. \end{cases}$$

The open ball of radius $\varepsilon = 2^{-r}$ centered at $c \in T^{\mathbb{Z}^2}$ is

$$B_\varepsilon(c) = \{e \in T^{\mathbb{Z}^2} \mid e(\vec{x}) = c(\vec{x}) \text{ for all } \|\vec{x}\| \leq r\},$$

where (and from now on) we denote $\|(x,y)\| = |x| + |y|$. These balls form a base of the topology.

More generally, for any finite domain $D \subseteq \mathbb{Z}^2$ and configuration $c \in T^{\mathbb{Z}^2}$ we define the *cylinder set*

$$\mathrm{Cyl}(c,D) = \{e \in T^{\mathbb{Z}^2} \mid e(\vec{x}) = c(\vec{x}) \text{ for all } \vec{x} \in D\}$$

that contains all those configurations that agree with c in domain D.

Note that for sufficiently large r we have $D \subseteq E$ where

$$E = \{\vec{x} \in \mathbb{Z}^d \mid \|\vec{x}\| \leq r\}.$$

Then

$$\mathrm{Cyl}(c,D) = \bigcup_{e \in \mathrm{Cyl}(c,D)} \mathrm{Cyl}(e,E),$$

so all cylinders are (finite) unions of open balls, and hence they are open in the topology. Balls form a base of the topology, so also cylinders form a base.

The complement of cylinder $\mathrm{Cyl}(c,D)$ is

$$\bigcup_{e \notin \mathrm{Cyl}(c,D)} \mathrm{Cyl}(e,D),$$

so all cylinders are also closed, hence clopen. Our space has a clopen base. Clopen sets are exactly finite unions of cylinders (homework).

Let us next show that a sequence of configurations $c_1, c_2, \ldots$ converges to $c \in T^{\mathbb{Z}^2}$ in this topology, if and only if it converges to c according to the definition of convergence in Section 4.2. First, suppose convergence to c in the topology, and let $\vec{n} \in \mathbb{Z}^2$ be arbitrary. Denote

$$U = \mathrm{Cyl}(c, \{\vec{n}\}).$$

Convergence to c implies that for all sufficiently large i holds $c_i \in U$, that is, $c_i(\vec{n}) = c(\vec{n})$. So the sequence converges to c according to the definition of Section 4.2.

Conversely, suppose converge to c as defined in Section 4.2. Let U be an open set that contains c. Because cylinders form a base, there is a finite $D \subseteq \mathbb{Z}^d$ such that $\mathrm{Cyl}(c, D) \subseteq U$. By the definition of convergence of $c_1, c_2, \ldots$ there is $k \in \mathbb{Z}$ such that $c_i \in \mathrm{Cyl}(c, D)$ for all $i > k$. This means that the sequence converges to c in the topology.

Now we immediately obtain the following corollaries of our earlier propositions:

Corollary 6.17 *The metric space $T^{\mathbb{Z}^2}$ is compact.*

Proof. Follows directly from Propositions 4.2 and 6.8. □

Based on the propositions in the previous section, every compact metric space is a Hausdorff, separable and second countable, so the space $T^{\mathbb{Z}^2}$ has all these properties.

Next we look into translations and show that they are continuous functions. A *translation* by $\vec{n} \in \mathbb{Z}^2$ is the transformation $\tau_{\vec{n}} : T^{\mathbb{Z}^2} \longrightarrow T^{\mathbb{Z}^2}$ that maps $c \mapsto e$ where $e(\vec{m}) = c(\vec{m} - \vec{n})$ for all $\vec{m} \in \mathbb{Z}^2$. Translations are bijective, and $\tau_{\vec{n}}$ and $\tau_{-\vec{n}}$ are inverses of each other. The *east shift* σ_e and the *north shift* σ_n are translations by vectors $(1, 0)$ and $(0, 1)$ respectively, and the *west* and the *south shifts* are their inverses $\sigma_w = \sigma_e^{-1}$ and $\sigma_s = \sigma_n^{-1}$. All translations are compositions of the four shifts. Let us denote by $\mathbb{T}$ the set of all translations.

For every $\vec{n} \in \mathbb{Z}^2$ and $D \subseteq \mathbb{Z}^2$ we denote the translation of D by $\vec{n}$ as

$$D + \vec{n} = \{\vec{d} + \vec{n} \mid \vec{d} \in D\}.$$

Let $\mathrm{Cyl}(c, D)$ be an arbitrary cylinder. Because

$$\tau_{\vec{n}}\left(\mathrm{Cyl}(c, D)\right) = \mathrm{Cyl}(\tau_{\vec{n}}(c), D + \vec{n})$$

we have that translations $\tau_{\vec{n}}$ are continuous.

So we have a compact, metric space $T^{\mathbb{Z}^2}$, equipped with continuous transformations generated by $\sigma_s, \sigma_e, \sigma_n, \sigma_w$. This is a set-up studied by topological dynamics.

6.3 Subshifts

A set $A \subseteq T^{\mathbb{Z}^2}$ is *translation invariant* if $\tau(A) = A$ for every $\tau \in \mathbb{T}$. For translation invariance it is enough to verify that $\sigma_e(A) = A$ and $\sigma_n(A) = A$. A topologically closed, translation invariant set is a (two-dimensional) *subshift*, while the entire configuration space $T^{\mathbb{Z}^2}$ is also called the (two-dimensional) *full shift* over the alphabet T.

Recall from the beginning of Section 5.6 that a finite pattern over T is a pair (D, p) where $D \subseteq \mathbb{Z}^2$ is finite, the domain of the pattern, and $p : D \longrightarrow T$. Let us denote by $\mathcal{P}(T)$ the set of all finite patterns over T. Clearly $\mathcal{P}(T)$ is countable as the number of finite subsets of $\mathbb{Z}^2$ is countable.

Pattern (D, p) is a *subpattern* of $c \in T^{\mathbb{Z}^2}$ if $c(\vec{n}) = p(\vec{n})$ for all $\vec{n} \in D$. Configurations that have (D, p) as a subpattern form a cylinder which we denote by

$$\mathrm{Cyl}(p, D) = \{c \in T^{\mathbb{Z}^2} \mid c(\vec{n}) = p(\vec{n}) \text{ for all } \vec{n} \in D\}.$$

This is of course the same cylinder as $\mathrm{Cyl}(c, D)$ for any configuration c in the cylinder.

We say that the pattern (D, p) *appears* in c if (D, p) is a subpattern of $\tau(c)$ for some translation $\tau \in \mathbb{T}$. For any configuration c let $\mathrm{Patt}(c)$ be the set of all finite patterns that appear in c, and for any $A \subseteq T^{\mathbb{Z}^2}$ we denote by

$$\mathrm{Patt}(A) = \bigcup_{c \in A} \mathrm{Patt}(c)$$

the set of finite patterns that appear in some element of A.

For any set P of finite patterns we define the set

$$\Sigma(P) = \{c \in T^{\mathbb{Z}^2} \mid \mathrm{Patt}(c) \cap P = \emptyset\}$$

of configurations in which no element of P appears. Next we prove that sets $\Sigma(P)$ are precisely the subshifts over T.

Theorem 6.18 *$\Sigma \subseteq T^{\mathbb{Z}^2}$ is a subshift if and only if $\Sigma = \Sigma(P)$ for some set P of finite patterns over T.*

Proof. First we observe that $\Sigma(P)$ is a subshift, for every $P \subseteq \mathcal{P}(T)$. Clearly $\Sigma(P)$ is translation invariant because $\mathrm{Patt}(c) = \mathrm{Patt}(\tau(c))$ for all configurations c and translations τ. And $\Sigma(P)$ is closed because for every $c \notin \Sigma(P)$ there exists $(D, p) \in P$ that appears in c, i.e., is a subpattern of $\tau(c)$ for some translation τ. Then $\tau(c) \in \mathrm{Cyl}(p, D)$ and $\mathrm{Cyl}(p, D) \cap \Sigma(P) = \emptyset$. This means that $\tau^{-1}(\mathrm{Cyl}(p, D))$ is an open neighborhood of c whose intersection with $\Sigma(P)$ is empty.

For the converse direction, let Σ be an arbitrary subshift over T and define

$$P = \mathcal{P}(T) \setminus \bigcup_{x \in \Sigma} \mathrm{Patt}(x),$$

that is, P contains all the patterns that do not appear in any configuration belonging to Σ. Let us prove that $\Sigma = \Sigma(P)$. If $c \in \Sigma$ then by the definition of P we have $\mathrm{Patt}(c) \cap P = \emptyset$. Hence $c \in \Sigma(P)$. And if $c \in \Sigma(P)$ then $\mathrm{Patt}(c) \subseteq \cup_{x \in \Sigma} \mathrm{Patt}(x)$, so for every finite $D \subseteq \mathbb{Z}^2$ we have $\Sigma \cap \mathrm{Cyl}(c, D) \neq \emptyset$. Because Σ is closed we have $c \in \Sigma$. □

Subshifts $\Sigma(P)$ for finite P are called *subshifts of finite type (SFT)*. So a SFT can be specified by giving a finite collection P of forbidden patterns. But this is exactly how we defined valid tilings in Section 5.3. So the valid tilings form a SFT, and conversely, every SFT is the set of valid tilings when the forbidden patterns are defined as in the SFT. Valid tilings by Wang tiles are particular types of subshifts of finite type, with small forbidden patterns. The construction in Lemma 5.5 in Section 5.3 in fact shows the following: Every subshift of finite type is conjugate to the set of valid tilings by some Wang tile set. (Two subshifts X and Y are called *conjugate* if there exists a translation commuting homeomorphism between them, i.e., a continuous bijection $h : X \longrightarrow Y$ such that $h \circ \tau = \tau \circ h$ for all $\tau \in \mathbb{T}$. Conjugate subshifts are in many respects equivalent with each other.)

6.4 Orbits, transitivity and minimality

For any $c \in T^{\mathbb{Z}^2}$ the set

$$\mathcal{O}(c) = \{\tau(c) \mid \tau \in \mathbb{T}\}$$

is the *orbit* of c. The set $\mathcal{O}(c)$ is trivially translation invariant. The orbit is not necessarily closed, so we frequently consider the *orbit closure* $\overline{\mathcal{O}(c)}$. The orbit closure is translation invariant: Let $e \in \overline{\mathcal{O}(c)}$ and let τ be any translation. Since there are $e_1, e_2, \cdots \in \mathcal{O}(c)$ such that $\lim_{i \longrightarrow \infty} e_i = e$, we have $\lim_{i \longrightarrow \infty} \tau(e_i) = \tau(e)$ and each $\tau(e_i) \in \mathcal{O}(c)$. This means that $\tau(e) \in \overline{\mathcal{O}(c)}$. We have proved the following:

Lemma 6.19 *The orbit closure $\overline{\mathcal{O}(c)}$ is a subshift, for every $c \in T^{\mathbb{Z}^2}$.* □

The orbit closure $\overline{\mathcal{O}(c)}$ is the subshift generated by c, that is, the intersection of all subshifts that contain c.

Example 21. Let $T = \{0, 1\}$ and let $c \in T^{\mathbb{Z}^2}$ be the infinite cross: $c(i, 0) = c(0, i) = 1$ for all $i \in \mathbb{Z}$ and $c(i, j) = 0$ if $i, j \neq 0$. Then $\mathcal{O}(c)$ is not closed. The orbit closure also contains the zero configuration c_0 with $c_0(i, j) = 0$ for all $i, j \in \mathbb{Z}$, as well as horizontal and vertical rows of 1's, i.e. elements of $\mathcal{O}(c_v)$ and $\mathcal{O}(c_h)$ where $c_h(i, 0) = c_v(0, i) = 1$ for all $i \in \mathbb{Z}$ and $c_h(i, j) = c_v(j, i) = 0$ if $j \neq 0$. □

Lemma 6.20 *$e \in \overline{\mathcal{O}(c)}$ if and only if Patt$(e) \subseteq$ Patt(c).*

Proof. We have $e \in \overline{\mathcal{O}(c)}$ if and only if for every finite $D \subseteq \mathbb{Z}^2$ there exists $\tau \in \mathbb{T}$ such that $\tau(c) \in \mathrm{Cyl}(e, D)$. But this is equivalent to $\mathrm{Patt}(e) \subseteq \mathrm{Patt}(c)$. □

A non-empty subshift Σ is called *transitive* if for every $(D_1, p_1), (D_2, p_2) \in \mathrm{Patt}(\Sigma)$ there exists $c \in \Sigma$ such that $(D_1, p_1), (D_2, p_2) \in \mathrm{Patt}(c)$. In other words, any two patterns that appear in some elements of Σ, appear in the same element of Σ. Next we show that transitive subshifts are exactly the orbit closures of configurations:

Theorem 6.21 *Subshift Σ is transitive if and only if $\Sigma = \overline{\mathcal{O}(c)}$ for some $c \in \Sigma$.*

Proof. For every configuration c the subshift $\overline{\mathcal{O}(c)}$ is transitive: By Lemma 6.20 we have the inclusion $\mathrm{Patt}(\overline{\mathcal{O}(c)}) \subseteq \mathrm{Patt}(c)$, so all patterns that appear in some elements of $\overline{\mathcal{O}(c)}$ appear in c, and hence Σ is transitive.

Conversely, suppose that Σ is transitive. By transitivity and translation invariance of Σ, if U and V are cylinders such that $U \cap \Sigma \neq \emptyset$ and $V \cap \Sigma \neq \emptyset$ then there is a translation τ such that $U \cap \tau(V) \cap \Sigma \neq \emptyset$. And since non-empty intersections of cylinders are cylinders, the set $U \cap \tau(V)$ is a cylinder.

Let $U_1, U_2, \ldots$ be all the cylinders such that $U_i \cap \Sigma \neq \emptyset$. By the observation above, there are translations $\tau_1, \tau_2, \ldots$ such that

$$V_n = \tau_1(U_1) \cap \tau_2(U_2) \cap \cdots \cap \tau_n(U_n) \cap \Sigma$$

is non-empty for every $n = 1, 2, \ldots$. Every V_n is closed and $V_1 \supseteq V_2 \supseteq V_3 \supseteq \ldots$. By compactness there exists c in their intersection. This c is in Σ and it contains the patterns in $\mathrm{Patt}(\Sigma)$. This means that $\Sigma = \overline{\mathcal{O}(c)}$. □

So we can observe that if Σ is transitive then some $c \in \Sigma$ contains all the finite patterns that appear in any elements of Σ.

Motivated by the previous theorem, we call an element $c \in \Sigma$ *transitive* in Σ if $\Sigma = \overline{\mathcal{O}(c)}$. By the theorem, a transitive element c exists exactly in those subshifts that are transitive, and in that case $\mathcal{O}(c)$ is a dense subset of Σ consisting of transitive elements.

Example 22. In Example 21 the infinite cross is transitive in its orbit closure. The orbit closure contains a non-transitive subset $\Sigma = \overline{\mathcal{O}(c_v)} \cup \overline{\mathcal{O}(c_h)}$ generated by the horizontal and vertical rows of 1's. □

A non-empty subshift Σ is called *minimal* if the only subshifts contained in Σ are $\emptyset$ and Σ.

Theorem 6.22 *Let Σ be a subshift. The following are equivalent:*

(i) Σ is minimal.

(ii) All elements of Σ are transitive in Σ.

(iii) Patt(e) = Patt(c) for all $e, c \in \Sigma$.

Proof. (i) $\implies$ (ii): For every $c \in \Sigma$ the orbit closure $\overline{\mathcal{O}(c)}$ is a subshift inside Σ, so by minimality $\overline{\mathcal{O}(c)} = \Sigma$.

(ii) $\implies$ (i): If Σ is not minimal then it properly contains a non-empty subshift $\Sigma' \subsetneq \Sigma$. If $c \in \Sigma'$ then $\overline{\mathcal{O}(c)} \subseteq \Sigma'$, so c is not transitive in Σ.

(ii) $\implies$ (iii): By the definition of transitivity, $\overline{\mathcal{O}(e)} = \Sigma = \overline{\mathcal{O}(c)}$ for all $e, c \in \Sigma$. By Lemma 6.20 this means that $\mathrm{Patt}(e) = \mathrm{Patt}(c)$.

(iii) $\implies$ (ii): If $\mathrm{Patt}(e) = \mathrm{Patt}(c)$ for all $e, c \in \Sigma$ then by Lemma 6.20 we have $e \in \overline{\mathcal{O}(c)}$. As $e \in \Sigma$ is arbitrary, we have $\Sigma \subseteq \overline{\mathcal{O}(c)}$, i.e., c is transitive in Σ. □

So the theorem states that Σ is minimal if and only if the orbits of all its elements are dense in Σ. Next we show that minimal subshifts are found inside all non-empty subshifts. This could be proved directly using Zorn's lemma. We present an elementary topological proof.

Theorem 6.23 *Every non-empty subshift Σ contains a minimal subshift.*

Proof. Cylinders form a countable base $U_1, U_2, \ldots$ of the topology. Let us denote by

$$\mathcal{O}(U_i) = \{\tau(c) \mid \tau \in \mathbb{T}, c \in U_i\} = \bigcup_{\tau \in \mathbb{T}} \tau(U_i)$$

the orbit U_i. It is clearly translation invariant, and also open as a union of open sets $\tau(U_i)$.

Inductively we construct a sequence $F_0 \supseteq F_1 \supseteq F_2 \supseteq \ldots$ of non-empty, closed, translation invariant sets as follows. $F_0 = \Sigma$. Then suppose that F_{m-1} has been defined. If $F_{m-1} \subseteq \mathcal{O}(U_m)$ then $F_m = F_{m-1}$, else $F_m = F_{m-1} \setminus \mathcal{O}(U_m)$. Then $F_m \neq \emptyset$, F_m is closed as F_{m-1} is closed and $\mathcal{O}(U_m)$ is open, and F_m is translation invariant because F_{m-1} and $\mathcal{O}(U_m)$ are translation invariant. Let

$$F = \bigcap_{i=1}^{\infty} F_i.$$

Because all F_i are closed and translation invariant, so is F, and it follows from the compactness that $F \neq \emptyset$. So F is a non-empty subshift.

Let us show that F minimal. Suppose on the contrary that there exist $e, c \in F$ such that $\mathrm{Patt}(e) \setminus \mathrm{Patt}(c) \neq \emptyset$. This means that there is a cylinder U_i such that $e \in \mathcal{O}(U_i)$ and $c \notin \mathcal{O}(U_i)$. As $F \subseteq F_{i-1}$ is not a subset of $\mathcal{O}(U_i)$, we see that $F_i = F_{i-1} \setminus \mathcal{O}(U_i)$. But this contradicts the assumption $e \in \mathcal{O}(U_i)$. □

6.5 Periodicity and recurrence properties

A configuration $c \in T^{\mathbb{Z}^2}$ is one-way periodic if there exists $\vec{n} \in \mathbb{Z}^2 \setminus \vec{0}$ such that $c = \tau_{\vec{n}}(c)$. Vector $\vec{n}$ is a period of c. Configuration c is two-way periodic if it is periodic with two linearly independent periods $\vec{n}_1$ and $\vec{n}_2$. A two-way periodic configuration is always periodic with horizontal and vertical periods $(0, n)$ and $(n, 0)$ for some $n > 0$, as shown in the beginning of Section 4.1. If a subshift of finite type contains a one-way periodic element then it contains a two-way periodic element as well: this was show for valid Wang tilings in Theorem 4.1, and by Lemma 5.5 subshifts of finite type are conjugate to sets of valid Wang tilings. Note that a conjugacy preserves vectors of periodicity.

The orbit $\mathcal{O}(c)$ of c is finite if and only if c is two-way periodic. The orbit is also closed if and only if c is two-way periodic (homework).

Two-way periodicity is a very strong form of recurrence. Some weaker recurrence properties are defined in the following. A configuration $c \in T^{\mathbb{Z}^2}$ is *uniformly recurrent* if for every open U with $c \in U$ there exists a finite $D \subseteq \mathbb{Z}^2$ such that for every $\vec{n} \in \mathbb{Z}^2$ we have $\tau_{\vec{n}+\vec{d}}(c) \in U$ for some $\vec{d} \in D$. In other words, if a finite pattern appears somewhere in a uniformly recurrent c then it appears inside every $n \times n$ square, for some n.

A configuration is called *recurrent* if for every open U with $c \in U$ there exists $\vec{n} \neq \vec{0}$ such that $\tau_{\vec{n}}(c) \in U$. In other words, every pattern that appears in c appears more than once. It is easy to see that then the pattern has to appear infinitely many times in c:

Lemma 6.24 *Configuration c is recurrent if and only if for every open neighborhood U of c there are infinitely many translations $\tau \in \mathbb{T}$ such that $\tau(c) \in U$.*

Proof. Let c be recurrent, and let U be open, $c \in U$. Suppose, contrary to the claim, that the only translations τ such that $\tau(c) \in U$ are by vectors $\vec{n}_1, \vec{n}_2, \ldots \vec{n}_k$. Let

$$V = \bigcap_{i=1}^{k} \tau_{-\vec{n}_i}(U).$$

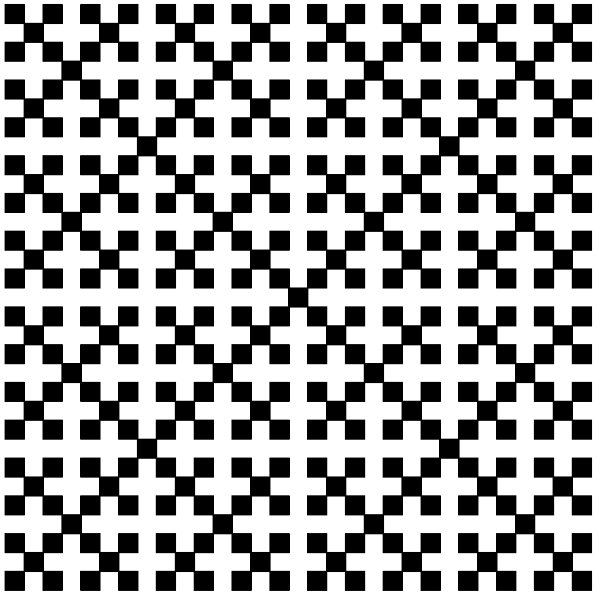
As V is open and $c \in V$, by the recurrence of c there exists $\vec{n} \neq \vec{0}$ such that $\tau_{\vec{n}}(c) \in V$. Then $\tau_{\vec{n}_i+\vec{n}}(c) \in U$ for all $i = 1, 2, \ldots, k$. This means that $\{\vec{n}_1 + \vec{n}, \vec{n}_2 + \vec{n}, \ldots \vec{n}_k + \vec{n}\} = \{\vec{n}_1, \vec{n}_2, \ldots \vec{n}_k\}$. This is possible only if $\vec{n} = \vec{0}$, a contradiction. (Namely, suppose w.l.o.g. that $\vec{n} = (x, y)$ where $x > 0$. If $\vec{n}_i$ has the largest x-coordinate among $\vec{n}_1, \vec{n}_2, \ldots \vec{n}_k$, then $\vec{n}_i + \vec{n}$ would have a larger x-coordinate than any of the vectors in the set.) □

Configuration c is *quasi-periodic* if for every open neighborhood U there exist linearly independent translation vectors $\vec{a}, \vec{b} \in \mathbb{Z}^2$ such that $\tau_{i\vec{a}+j\vec{b}}(c) \in U$ for all $i, j \in \mathbb{Z}$. In other words, in quasi-periodic configurations all finite patterns are part of a two-way periodic repetition of the pattern, but the period may be different for different patterns. Configuration c is called *isochronous* if for every open neighborhood U there exists an offset vector $\vec{c} \in \mathbb{Z}^2$ and two linearly independent $\vec{a}, \vec{b} \in \mathbb{Z}^2$ such that $\tau_{i\vec{a}+j\vec{b}+\vec{c}}(c) \in U$ for all $i, j \in \mathbb{Z}$.

The following implications are obvious from the definitions:

$$c \text{ two-way periodic} \implies c \text{ quasi-periodic} \implies c \text{ isochronous} \implies c \text{ uniformly recurrent} \implies c \text{ recurrent}$$

Example 23. For any integer n let us define $\deg_2(n) = k$ if $n = a2^k$ for some odd a, and $\deg_2(0) = \infty$. Let $T = \{0, 1\}$. Define configuration $c \in T^{\mathbb{Z}^2}$ as follows: $c(i, j) = 1$ if and only if $\deg_2(i) = \deg_2(j)$. The following figure illustrates c where black square indicates value 1 and white square value 0:



This c is isochronous but not quasi-periodic, as defined above. The configuration is not quasi-periodic, because symbol 1 in position $(0,0)$ is not repeated two-way periodically: all other cells at $(i,0)$ and $(0,i)$ carry symbol 0. But the configuration is isochronous (and hence uniformly recurrent and recurrent) because $\deg_2(n) = \deg_2(n+2^k)$ for every $k > \deg_2(n)$. This means that for any odd integers a and b the translated configuration $\tau_{(a2^k,b2^k)}(c)$ agrees with c inside the square $\{(i,j) \in \mathbb{Z}^2 \mid -2^k < i,j < 2^k\}$. □

Uniformly recurrent configurations are important because they exactly generate all minimal subshifts.

Theorem 6.25 *Subshift $\overline{\mathcal{O}(c)}$ is minimal if and only if c is uniformly recurrent.*

Proof. By Theorem 6.22 subshift $\overline{\mathcal{O}(c)}$ is minimal if and only if $\mathrm{Patt}(e) = \mathrm{Patt}(c)$ for all $e \in \overline{\mathcal{O}(c)}$.

"$\Longleftarrow$": Let c be uniformly recurrent and let $e \in \overline{\mathcal{O}(c)}$ arbitrary. By Lemma 6.20 we know that $\mathrm{Patt}(e) \subseteq \mathrm{Patt}(c)$ so it is enough to show that $\mathrm{Patt}(c) \subseteq \mathrm{Patt}(e)$. Let $(D,p) \in \mathrm{Patt}(c)$. By uniform recurrence of c, there exists n such that every $n \times n$ square in c contains a copy of (D,p). Because every $n \times n$ square pattern in e also appears in c, every $n \times n$ square pattern of e contains (D,p). So $(D,p) \in \mathrm{Patt}(e)$.

"$\Longrightarrow$": Conversely, assume that c is not uniformly recurrent. Let us show that $\Sigma = \overline{\mathcal{O}(c)}$ is not minimal. By uniform recurrence there exists a finite domain $D \subseteq \mathbb{Z}^2$ such that for any finite $\mathbb{F} \subseteq \mathbb{T}$ a translation $\alpha \in \mathbb{T}$ exists such that for all $\tau \in \mathbb{F}$ holds $\alpha\tau(c) \notin \mathrm{cyl(c,D)}$. In particular, if $\mathbb{T} = \{\tau_1, \tau_2, \dots\}$ then for every $j = 1, 2, \dots$ there exists $\alpha_j \in \mathbb{T}$ such that $\alpha_j\tau_i(c) \notin \mathrm{cyl(c,D)}$ for all $i = 1, 2 \dots, j$. Let e be the limit of a converging subsequence of $\alpha_1(c), \alpha_2(c), \alpha_3(c), \dots$. Then $e \in \Sigma$ but for all $\tau_i \in \mathbb{T}$ holds $\tau_i(e) \notin \mathrm{cyl(c,D)}$. This is due to the fact that for all $j \geq i$ we have $\tau_i\alpha_j(c) = \alpha_j\tau_i(c) \notin \mathrm{cyl(c,D)}$. We conclude that $\mathrm{Patt}(c) \setminus \mathrm{Patt}(e) \neq \emptyset$, so Σ is not minimal. □

We have shown that a subshift Σ is minimal if and only if $\Sigma = \overline{\mathcal{O}(c)}$ for some uniformly recurrent c, and that in this case all elements of Σ are uniformly recurrent. There are two possibilities for the elements of a minimal Σ: Either all of them are two-way periodic, in which case the subshift is their finite orbit, or all elements are non-periodic (but uniformly recurrent) configurations that contain exactly the same finite patterns. In the second case the subshift turns out to contain an uncountable number of elements:

Theorem 6.26 *A minimal subshift is either finite or uncountably infinite.*

Proof. Let Σ be a minimal subshift of countable cardinality. By Proposition 6.13, there is an isolated point $c \in \Sigma$, so that some open set U satisfies $\Sigma \cap U = \{c\}$. By Theorem 6.25 configuration c is uniformly recurrent. This means that there are many translations τ such that $\tau(c) \in U$. In fact, for some n, every $n \times n$ square must contain a point $\vec{n}$ such that $\tau = \tau_{\vec{n}}$ has this property. Whenever $\tau_{\vec{n}}(c) \in U$ we have $\tau_{\vec{n}}(c) = c$, so c is periodic. This implies that $\Sigma = \overline{\mathcal{O}(c)}$ is finite. □

The following corollary states some implications of the above theorems in the case the subshift considered is the set of valid tilings:

Corollary 6.27 *If a tile set admits a valid tiling then it admits a uniformly recurrent tiling. If it admits a uniformly recurrent tiling that is not two-way periodic, then it admits uncountably many different tilings. In particular, every aperiodic tile set admits uncountably many valid, uniformly recurrent tilings.* □

6.6 Equicontinuity and isolated points

Recall that point $c \in \Sigma$ is *isolated* in Σ if there exists an open set U such that $U \cap \Sigma = \{c\}$.

Lemma 6.28 *Let Σ be a subshift. All $c \in \Sigma$ are isolated in Σ if and only if Σ is finite.*

Proof. Let Σ be a finite subshift, and let $c \in \Sigma$. Set $F = \Sigma \setminus \{c\}$ is closed as a finite union of singleton sets. Hence the complement of F is an open neighborhood of c that does not contain any other elements of Σ.

Conversely, suppose that Σ is an infinite subshift. By compactness there exists an infinite converging sequence $c_1, c_2, \ldots$ where each $c_i \in \Sigma$ and and $c_i \neq c_j$ whenever $i \neq j$. The limit $c = \lim_{i\to\infty} c_i$ is in Σ, but it is not isolated in Σ. □

Finite subshifts are exactly the ones whose elements are all two-way periodic:

Theorem 6.29 *A subshift Σ is finite if and only if every $c \in \Sigma$ is two-way periodic.*

Proof. If c is not two-way periodic then its orbit is infinite, so one direction is trivial. We only need to show that if all $c \in \Sigma$ are two-way periodic then Σ is finite.

Suppose the contrary: Σ is infinite and all $c \in \Sigma$ are two-way periodic. Due to infinity, there exists a converging sequence $c_1, c_2, \ldots$ such that all $c_i \in \Sigma$ and $c_i \neq c_j$ for all $i \neq j$. Let $c = \lim_{i\to\infty} c_i$. Because all elements of Σ are two-way periodic, $c \in \Sigma$ is two-way periodic.

For each $c_i \neq c$ let us identify a vector $(x_i, y_i) \in \mathbb{Z}^2$ of minimum $n_i = \max\{|x_i|, |y_i|\}$ such that $c_i(x_i, y_i) \neq c(x_i, y_i)$. In other words, $c_i(x, y) = c(x, y)$ for all $-n_i < x, y < n_i$, but $c_i(x_i, y_i) \neq c(x_i, y_i)$ for some x_i, y_i satisfying $|x_i| = n_i$ or $|y_i| = n_i$. Note that $\lim_{i\to\infty} n_i = \infty$.

Infinitely many of the vectors (x_i, y_i) are in the same quadrant of the plane. Without loss of generality we assume now that all $x_i, y_i \geq 0$. By setting $\tau_i = \tau_{(-x_i,-y_i)}$ we see that for all $i = 1, 2, \ldots$ $\tau_i(c_i)(\vec{0}) \neq \tau_i(c)(\vec{0})$ but $\tau_i(c_i)(x, y) = \tau_i(c)(x, y)$ for all $-n_i < x < 0$ and $-n_i < y < 0$. Because c is two-way periodic, there are only finitely many different configurations among $\tau_i(c)$. By choosing a subsequence, we can hence assume now that $\tau_i(c) = \tau(c)$ for some translation τ and all $i = 1, 2, \ldots$.

Sequence $\tau_1(c_1), \tau_2(c_2), \ldots$ has a converging subsequence. The limit $e \in \Sigma$ of the subsequence coincides with a two-way periodic configuration $\tau(c)$ at (x, y) for all $x, y < 0$, but it does not coincide with $\tau(c)$ at $(0, 0)$. All elements of Σ are two-way periodic, so e is two-way periodic. Configurations e and $\tau(c)$ have a common period (a, b) with $a, b < 0$, which implies that

$$\tau(c)(0, 0) = \tau(c)(a, b) = e(a, b) = e(0, 0),$$

a contradiction. □

The following term comes from topological dynamics: Configuration $c \in \Sigma$ is an *equicontinuity point* if

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall e \in \Sigma)(\forall \tau \in \mathbb{T})\ d(c,e) < \delta \Longrightarrow d(\tau(c), \tau(e)) < \varepsilon.$$

In other words, if e is chosen sufficiently close to c then all translates $\tau(e)$ and $\tau(c)$ are close to each other.

Because for any $c \neq e$ there exists $\tau \in \mathbb{T}$ such that $d(\tau(c), \tau(e)) = 1$, we see that c is an equicontinuity point in Σ if and only if it is isolated in Σ. Indeed, if c is isolated then for some $\delta > 0$ the only configuration e satisfying $d(c,e) < \delta$ is c itself. And conversely, if c is not isolated then all neighborhoods of c contain $e \neq c$, so the choice $\varepsilon = 1/2$ contradicts the equicontinuity condition at c.

The subshift Σ is called *equicontinuous* if all $c \in \Sigma$ are equicontinuity points. By the observation above, Σ is equicontinuous if and only if all its elements are isolated. By Lemma 6.28 equicontinuous subshifts are exactly the finite subshifts, which by Lemma 6.29 are exactly the subshifts all of whose elements are two-way periodic.

A subshift Σ is called *sensitive* if there exists $\varepsilon > 0$, called the sensitivity constant, such that

$$(\forall c \in \Sigma)(\forall \delta > 0)(\exists e \in \Sigma)(\exists \tau \in \mathbb{T})\ 0 < d(c,e) < \delta \text{ and } d(\tau(c), \tau(e)) > \varepsilon.$$

In other words, arbitrarily close to each c there is another configuration e such that for a suitable translation τ the configurations $\tau(c)$ and $\tau(e)$ are not close to each other. Note that if c is isolated in Σ then there are no elements within distance δ of c for small δ, and hence the system is not sensitive. In contrast, if there are no isolated points then the system is sensitive with any sensitivity constant $0 < \varepsilon < 1$ because, as pointed out above, for any $c \neq e$ we have $d(\tau(c), \tau(e)) = 1$ for a suitable $\tau \in \mathbb{T}$.

Finally, the fact that

$$(\exists \varepsilon > 0)(\forall c, e \in \Sigma)\ c \neq e \Longrightarrow\ (\exists \tau \in \mathbb{T})\ d(\tau(c), \tau(e)) > \varepsilon$$

means, in terms of topological dynamics terminology that all subshifts are *expansive*.

Theorem 6.30 *Let Σ be a subshift.*

(i) Σ is expansive.

(ii) Σ is sensitive if and only if it has no isolated points.

(ii) Σ is equicontinuous if and only all its elements are isolated, i.e., the subshift is finite.

□

7 A brief revisit to tilings by polygons

In the beginning of Section 4 we showed how any Wang protoset can be converted into an equivalent set of prototiles that are polygons, by replacing colors with suitable bumps and dents. By "equivalent" we mean that for every valid tiling by the polygons there is an isometry α that maps the tiling into another tiling where the tiles are aligned on integer lattice points so that the tiles – if replaced by the corresponding Wang tiles – provide a valid Wang tiling.

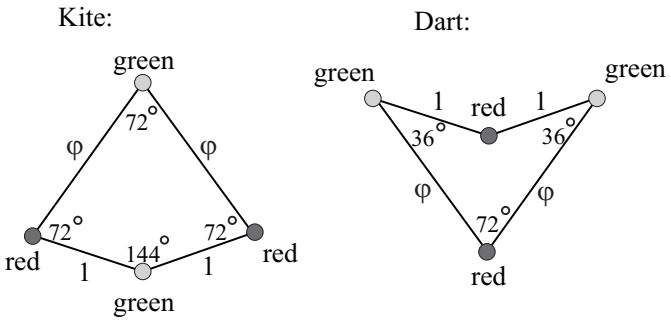
Using this construction of bumps and dents we can lift many results of the previous chapter to the case of polygonal prototiles. In particular, by Theorem 4.5 we know that there are aperiodic protosets of polygons, that is, finite sets of polygons that admit valid tilings but none of these tilings have translational symmetry. The bumps and dents prevent any reflectional or rotational symmetries so it is clear that the protosets we obtain only admit tilings without any non-trivial symmetries.

Theorem 7.1 *There exists a protoset of polygons that admits a valid tiling but does not admit a valid tiling with a non-trivial symmetry.*

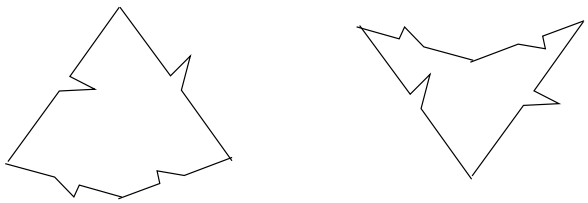
□

The smallest aperiodic Wang protoset contains 11 tiles, but with geometric tiles a single tile is enough to force non-periodicity! A polygonal prototile (the *hat*) was recently reported that is alone aperiodic: there exist monohedral tilings of $\mathbb{R}^2$ using the hat but none of these tilings has a translational symmetry. We discuss this tile in Section 7.4 below.

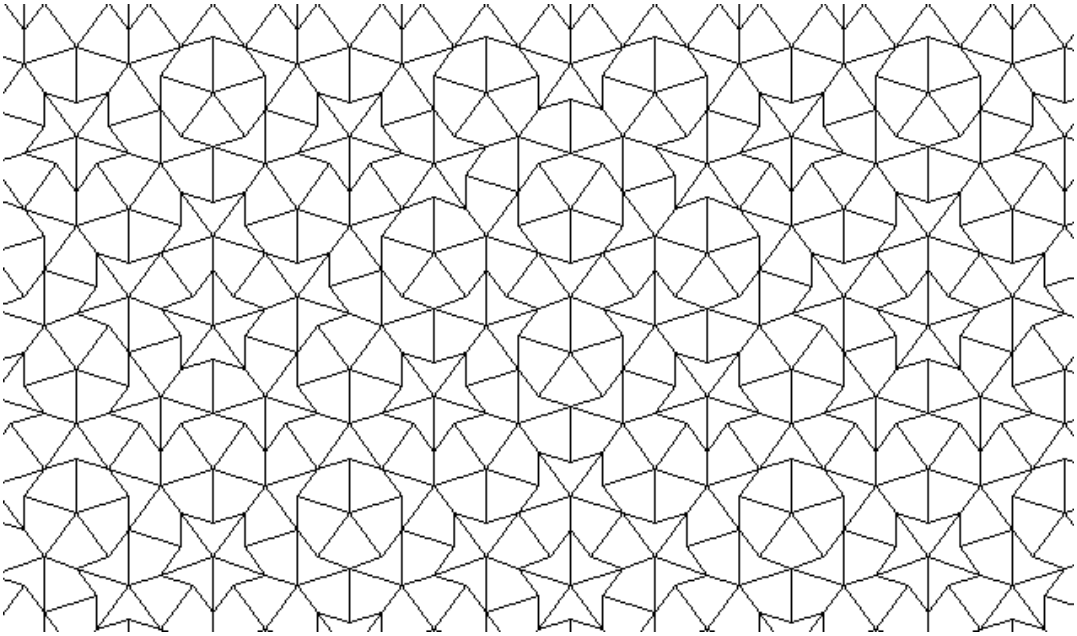
An older aperiodic protoset consists of two polygons. In 1974 R.Penrose presented this famous aperiodic pair called *kite* and *dart*:



The Penrose tiles are obtained by cutting in two a rhombus that has a 72° angle. The resulting quadrilaterals have edges of length 1 and $\varphi = (1 + \sqrt{5})/2 = 1.618\ldots$, the golden ratio. The vertices are colored red and green, and in valid tilings equal edges must be placed together and also the colors at the vertices must match. (This condition prevents one from gluing the kite and dart back together to form the rhombus.) These matching rules can be easily enforced using geometric constraints only by, say, using bumps and dents as follows:



Here is a part of a tiling using kites and darts. (For clarity, the bumps and the dents are not shown):



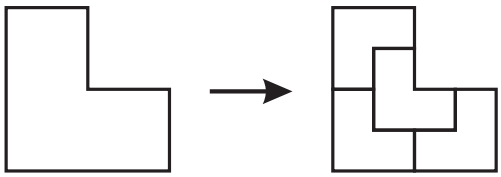
The following result will be proved in the homework problems:

Theorem 7.2 *Penrose kite and dart are an aperiodic pair of prototiles. They do admit valid tilings with a 5-fold rotational symmetry.*

□

7.1 Substitutions

Substitutions are a popular method to construct hierarchical, non-periodic tilings. As a simple example, consider the *chair substitution* where an L-tromino is cut into four smaller, similar shapes:

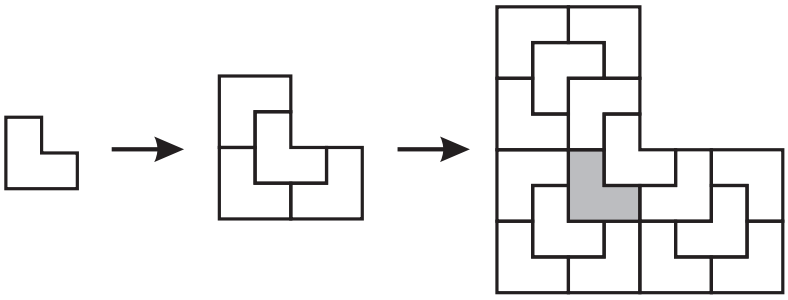


Starting from a single tile, we repeat the following operations:

(i) Replace each tile by four smaller copies as above,

(ii) Magnify the obtained pattern by factor two horizontally and vertically.

Here are the first two iterations (the meaning of the grey tile will be explained later):



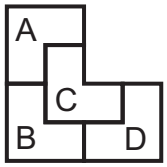
The first and the second steps produce "supertiles" that consist of four and sixteen copies of the tile. The k'th supertile is a pattern of 4^k tiles. It is clear that in this way we obtain ever larger areas covered by the tiles. To obtain a tiling of the infinite plane, we suitably position the obtained supertiles on the plane so that the next pattern always expands the previous one, and take the limit of the process. Note that the k'th supertile consists of four copies of the $(k-1)$'st supertile, so we can position it (in four different ways) on the plane so that the $(k-1)$'st supertile is its subpattern. Moreover, we can do this positioning so that the patterns grow in all directions, so that each point of the plane gets eventually covered by a tile.

For example, in the illustration above, we can align the gray tile of the second supertile over the initial tile, and repeat this positioning after all even rounds. Because the grey tile is not on the boundary, it is guaranteed to get surrounded by more and more tiles on all sides. The following illustration shows the position of the grey tile in the fourth supertile.

In the limit we obtain a tiling t of the infinite plane. It is clear that in this tiling each tile belongs to a supertile, which in turn belongs to a second level supertile, and so on.

To prove that the obtained tiling t is not periodic, we make the important observation that two supertiles cannot overlap: each tile in the tiling is part of a *unique* supertile. To see this, consider the

four parts of a supertile:



We observe that neighbors A and B meet at the end of the L, that is, they are adjacent to each other at an edge denoted by $*$ here:

Also neighbors B and D meet each other at their $*$-ends. In contrast, the $*$-ends of tile C are not adjacent to $*$-ends of other tiles. So we can recognize the center tiles C of all supertiles simply by the property that neither $*$-end is adjacent to a $*$-end of another tile. In the other cases A, B and D, the neighbor at the inner corner (indicated by $\bullet$ above) is the center tile C of the supertile, and hence the supertile containing the tile is unique also in this case.

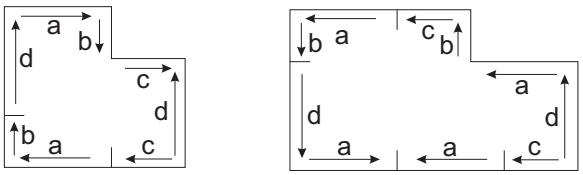
We conclude that tiling t has a unique partitioning into a tiling by the supertiles. The same reasoning then applies to the next levels, so that tiling t can be partitioned in a unique way into a tiling by k'th level supertiles, for every k.

Suppose now that tiling t would have a period $\vec{n} \neq (0,0)$. Applying translation $\tau_{\vec{n}}$ to the supertiling produces a supertiling. But since t remains the same under $\tau_{\vec{n}}$, and since the coresponding supertiling is unique, the supertiling must have period $\vec{n}$. The same reasoning applies to supertilings of level k, for every k. But for sufficiently large k, the k'th level supertile overlaps its translate by $\vec{n}$, so $\vec{n}$ cannot be a period of the k'th level supertiling, a contradiction.

We have proved that the tiling obtained by iterating the chair substitution is non-periodic. It was essential in the proof that the partitioning of the tiling into supertiles is unique. Note that the L-tromino is of course not aperiodic: it trivially tiles a 2×3 rectangle, which yields a periodic tiling of the plane. However, there are general methods to decorate tiles in such substitutions so that non-periodicity is forced. This increases the total number of tiles as several variants of the tiles are needed with different decorations. We skip the general method, but discuss in detail Amman's aperiodic tile set that is also based on the substitution concept.

7.2 Amman's aperiodic tile set

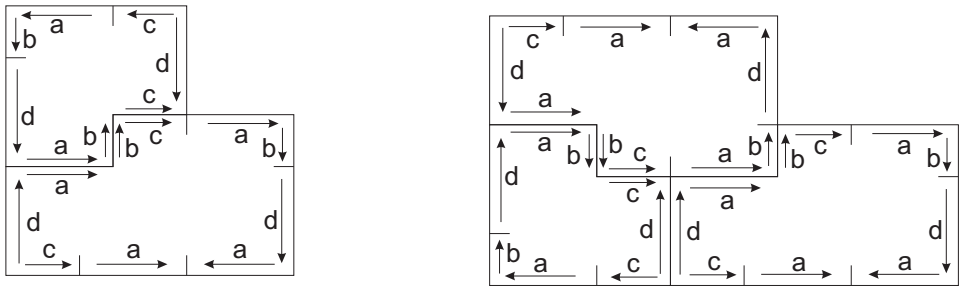
The following pair of tiles, due to R.Amman in 1977, forms an aperiodic tile set.



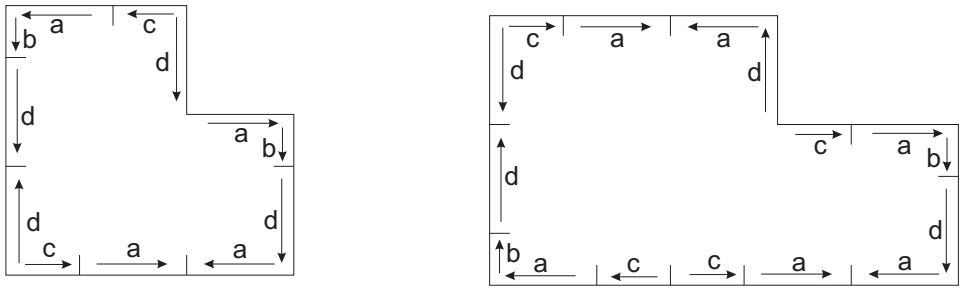
The tiles may be rotated and flipped in any orientation. The labeled arrows along the edges give a matching rule: each arrow must fit against an arrow with the same label and oriented in the same direction. The lengths of the arrows are arbitrary (yet positive), but all arrows with the same label have also the same length. (A remark: this matching rule can not be implemented geometrically with

bumps and dents. It can be implemented geometrically, though, by adding a third key tile, so that the construction provides an aperiodic set of three geometric tiles.)

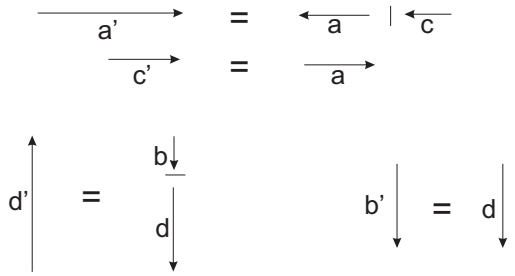
The following illustration shows how an A-tile and a B-tile fit together into a *super-A*, and how an A-tile and two B-tiles form a *super-B*:



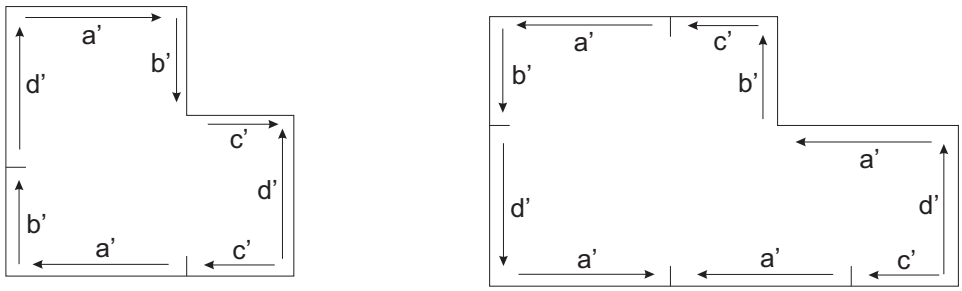
Any tiling by the resulting supertiles



can hence be broken into a tiling by the original tiles. Let us redecorate the supertiles with arrows labeled a', b', c' and d', where the new arrows represent combinations of old arrows as follows:



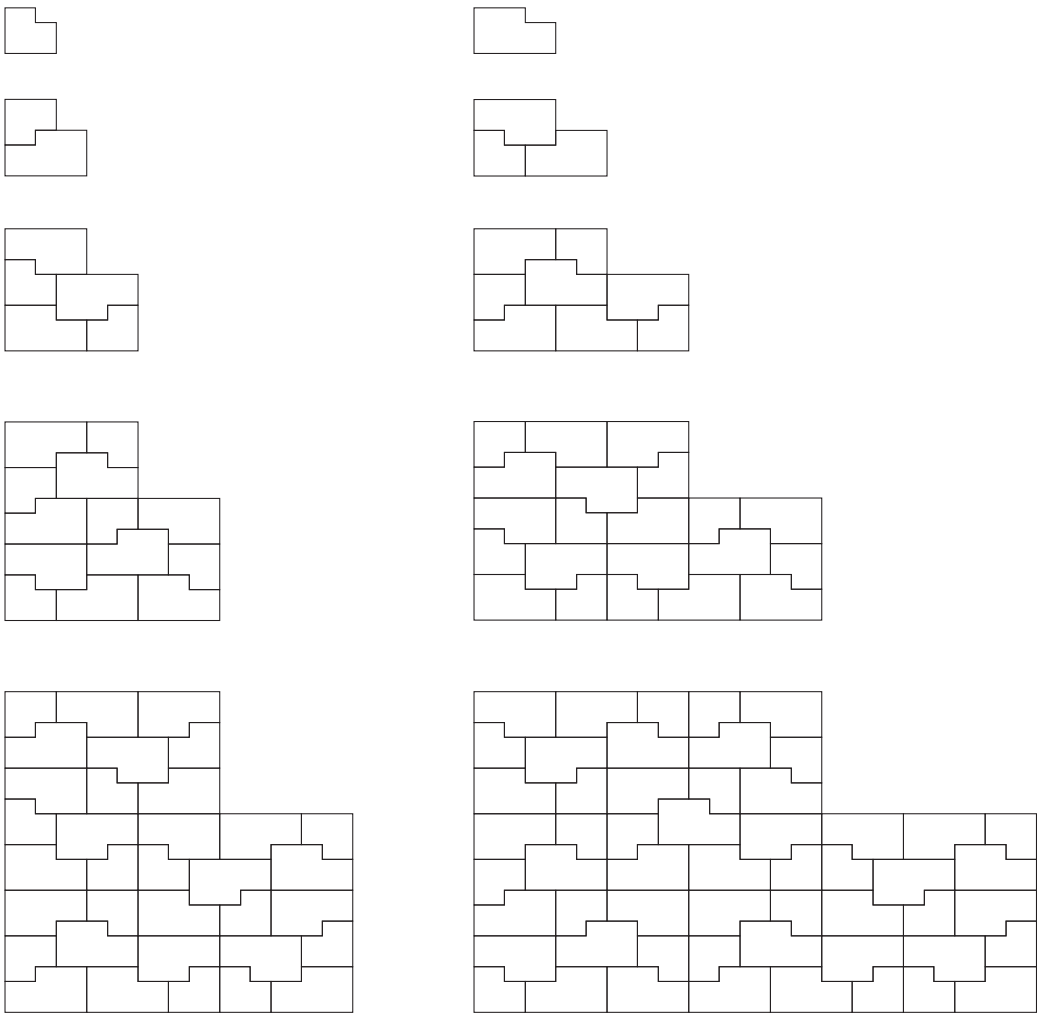
The redecorated supertiles



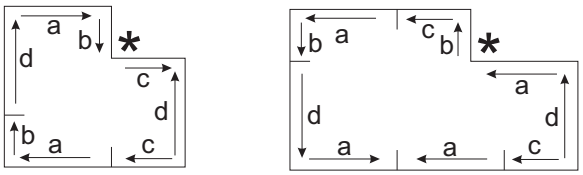
are called *expanded A* and *expanded B*, respectively. Of course, in any valid tiling by the expanded tiles, the tiles can be replaced by the corresponding supertiles, and the tiling remains valid. But also the converse holds: in any tiling by the supertiles, replacing the supertiles by the coresponding expanded tiles yields a valid tiling. To see this, note that in the supertiles the c-arrow is always immediately followed by an a-arrow in the same direction. Hence two neighbors with a common c-segment also share the following a-segment. So replacing every ca -segment by a new a'-arrow leaves the tiling valid. Analogously, every b-arrow is immediately followed by a d-arrow, so replacing bd -segment by a new d'-arrow keeps the tiling valid. But the obtained tiles are (up to renaming the arrows) exactly the expanded A- and B-tiles. We conclude that the supertiles and the expanded tiles admit exactly the same tilings.

Next we observe that the expanded tiles are isomorphic to the original tiles, where the arrows with labels a', b', c' and d' correspond to the arrows a, b, c and d, respectively. (However, the ratios of the arrow lengths may change, so the shapes of the expanded tiles are not necessarily similar to the original tiles. Similarity in shapes is obtained if the length of arrow d is φ times the length of b, and the length of a is φ times the length of c, where φ is the golden ratio, i.e., the positive number satisfying $\varphi^2 = \varphi + 1$. But similarity in shapes is not necessary in the reasoning that follows.)

We can now build supertiles of level two by simply combining the expanded tiles the same way we combined the original tiles to build the first level supertiles. Iterating the reasoning allows us to build supertiles and expanded tiles of levels two, three, four and so on. These provide tilings of larger and larger regions of the plane by the original A- and B-tiles. As with the chair substitution, we can take the limit, which yields a valid, hierarchical tiling of the infinite plane. The following illustration shows the first levels of the process:

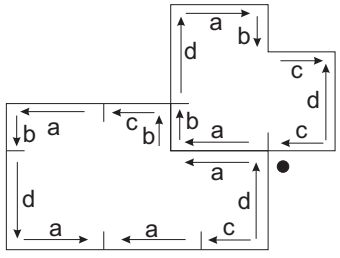


Consider now an arbitrary tiling of the plane by A and B. First we prove that every tile belongs to a supertile. Consider an A-tile and its neighbor at the inner corner *.

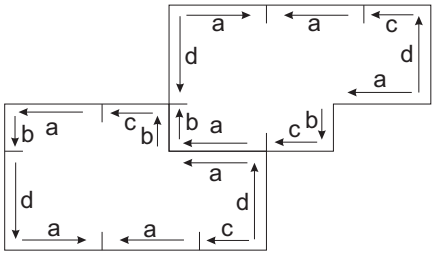


A simple case analysis (based, for example, on the b-arrows) shows that the only tile to match is the B-tile, oriented to form the super-A.

Consider then a B-tile and its inner corner *. An A-tile fits in the corner in two different ways: one provides super-A, as seen above; the other (shown below) cannot be expanded into a tiling of the plane, because no tile fits in the corner •:



On the other hand, there is only one possible way to fit a B-tile in the *-corner of the B-tile:



The only tile to fit in the *-corner of the second B-tile is the A-tile, so that the three tiles together form super-B. We conclude that every tile of a valid tiling is part of a super-A or a super-B.

In fact, the tiles of any tiling by A- and B-tiles can be grouped into non-overlapping supertiles and, moreover, such grouping is unique. To see this, find first all B-tiles whose *-neighbor is also a B-tile. These necessarily are part of a super-B, as discussed above. The super-B tiles do not overlap. All remaining tiles must be part of super-A tiles, so the unique grouping is concluded.

We are ready to make the following conclusion:

Theorem 7.3 *The A- and B-tiles form an aperiodic pair of tiles.*

Proof. By iterating the substitution to form supertiles, while properly aligning the obtained supertiles, we obtain a sequence of growing, correctly tiled patterns. In the limit, a valid tiling is obtained.

Let us show that no periodic tiling is possible. Suppose the contrary: A valid tiling t exists that is invariant under the translation by $\vec{n} \neq \vec{0}$. The tiles in t can be partitioned in a unique way into supertiles. If this tiling t_s by supertiles is translated by vector $\vec{n}$, a tiling t'_s by the supertiles is obtained. However, when the supertiles in t'_s are broken into their A- and B-pieces, the obtained tiling is the $\vec{n}$-translation of t, hence it is equal to t. But the supertiling obtained from t is unique, so $t'_s = t_s$, and tiling t_s is invariant under the translation by $\vec{n}$.

When the supertiles in t_s are replaced by the corresponding extended tiles, a valid tiling t_e by the extended tiles is obtained that has period $\vec{n}$. By repeating this argument on t_e in place of t, and iterating the reasoning, we see that there are valid tilings by extended tiles of all levels that are $\vec{n}$-periodic. This is not possible since an extended tile of a sufficiently high level overlaps with its translation by $\vec{n}$.

We conclude that the A- and B-tiles do not admit a periodic tiling, and hence they are an aperiodic pair of tiles. □

7.3 The extension and the periodicity theorems

To prove that repeating a substitution leads, in the limit, to a tiling of the plane was easy in our examples. We used the fact that the pattern obtained at iteration $k+1$ contains the level k pattern as its subpattern. This means that the sequence of obtained patterns have a well-defined limit. Moreover, in our examples the patterns grow on all sides, so that the limit covers the whole plane and is thus a valid tiling of the plane.

But it can be shown that even if the obtained patterns do not contain previous ones as subpatterns, a valid tiling exists as long as arbitrarily large disks can be covered. This result (which we present without a proof) extends Corollary 4.3 from Wang tiles to geometric tiles. It is important to note that the theorem only considers finite protosets of topological disks:

Theorem 7.4 (Extension theorem) *A finite protoset $\mathcal{P}$ of topological disks admits a tiling if and only if, for every $r > 0$, a disk of radius r can be covered by copies of the prototiles. (That is, there is a collection of tiles, all congruent to elements of $\mathcal{P}$, such that (i) the interiors of the tiles are pairwise disjoint, and (ii) a disk of radius r is included in the union of the tiles.)*

□

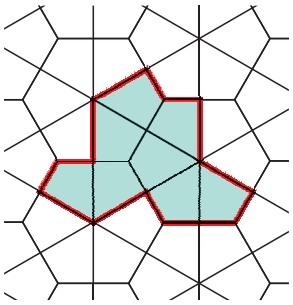
Theorem 7.4 above generalizes Corollary 4.3 from Wang tiles to geometric tiles. The other important basic property of Wang tiles states that a Wang tile set that admits a one-way periodic tiling admits also a two-way periodic tiling (Theorem 4.1). Also this theorem can be generalized to finite sets of polygonal prototiles and edge-to-edge tilings:

Theorem 7.5 (Periodicity theorem) *Let $\mathcal{P}$ be a finite set of polygons. Assume that there exists an edge-to-edge tiling by the protoset $\mathcal{P}$ that is one-way periodic (=invariant under some translation). Then there also exists an edge-to-edge tiling by $\mathcal{P}$ that is two-way periodic (=invariant under translations by two linearly independent vectors).*

□

7.4 Hat: an aperiodic monotile

The polygonal tile *hat* was recently proved to be an aperiodic monotile, *i.e.*, there exist monohedral tilings of the plane using the hat, but none of these tilings are periodic. The tile is a union of eight *kites* of a grid formed by overlapping the regular tilings by equilateral triangles and regular hexagons:



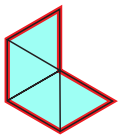
The grid of equilateral triangles is formed by equally spaced infinite lines in three directions, at 120° angles with each other. We number these directions 1, 2 and 3. For each direction $i \in \{1, 2, 3\}$ we call *i-lines* the infinite lines in direction i that together form the triangular lattice.

The grid of kites is in fact the dual of the $3 \cdot 4 \cdot 6 \cdot 4$ Archimedean tiling, obtained by joining the centers of adjacent tiles of the Archimedean tiling. A kite in the grid is a quadrilateral with two short edges (length 1) and two long edges (length $\sqrt{3}$). The long edges are along the i-lines, while the short edges are on the boundaries of the regular hexagons. The hat tile inherits these edge lengths. Note that when viewed as a polygon the hat has also one edge of length 2 (formed by two consecutive parallel short edges of kites), but in the following we consider this segment to consist of two edges of length 1. Thus the hat has 6 long edges and 8 short edges.

Let us read the edges as vectors, moving around the hat clockwise. We get 6 vectors of length $\sqrt{3}$, in three pairs of opposite vectors, and 8 vectors of length 1. The sum of the vectors is the zero vector (as the sum represents the vector leading from a vertex to itself). In fact, both the short and the long vectors separately sum up to zero:

- (H1) The sum of the six long vectors on the boundary of the hat is the zero vector, as is the sum of the eight short vectors.

Based on (H1) we can deform the hat tile as follows: keeping the orientations of the edges unchanged, scale the lengths of all long edges by some constant $a \geq 0$ and the lengths of all short edges by some constant $b \geq 0$, with $a \neq 0$ or $b \neq 0$. The scaled long edge vectors naturally still sum up to zero, and the scaled short edge vectors sum up to zero as well. The scaled edges define a boundary of a deformed tile. In our proof below we actually only need the scaling factors $a = 1$ and $b = 0$, *i.e.*, we remove the short edges. This results in the deformed tile that is the *chevron*



This is a hexagon with three pairs of equally long and parallel sides. It is also a union of four equilateral triangles with sides $\sqrt{3}$. On the other hand, three kites tile an equilateral triangle with side $2\sqrt{3}$. Thus the area of the chevron is the same as the area of three kites. Because the hat consists of eight kites, we see that

- (H2) the area of the hat is $8/3$ times the area of the chevron.

Let us argue next that deforming the tiles as above in a valid hat tiling produces a corresponding valid tiling by the deformed tiles. So let $\mathcal{T}$ be a valid monohedral tiling using the hat. Let $\mathcal{V} \subseteq \mathbb{R}^2$ be the set of vertices of $\mathcal{T}$, and let us assume, without loss of generality, that $\vec{0} \in \mathcal{V}$. Let us determine how the tile deformation moves an arbitrary vertex $\vec{v} \in \mathcal{V}$.

A *path* p is a sequence $\vec{v}_0, \vec{v}_1, \ldots, \vec{v}_n \in \mathcal{V}$ of vertices such that for each $i \in \{1, 2, \ldots n\}$ some tile in $\mathcal{T}$ has an edge from $\vec{v}_{i-1}$ to $\vec{v}_i$. The path thus tracks edges of the tiles. Denote $\vec{a}_i = \vec{v}_i - \vec{v}_{i-1}$, and let $L \subseteq \{1, 2, \ldots n\}$ be the set of indices such that $\vec{a}_i$ is long (of length $\sqrt{3}$, that is), and let $S = \{1, 2, \ldots n\} \setminus L$ be the index set of the short edges (of length 1). If $\vec{v}_n = \vec{v}_0$ then the path is a *cycle*. If, moreover, $\vec{v}_i \neq \vec{v}_j$ for all $i \neq j$ except when $\{i, j\} = \{0, n\}$ then the cycle is *simple*. For each cycle it is clear that

$$\sum_{i=1}^{n} \vec{a}_i = \vec{0},$$

but in fact we also have the property that

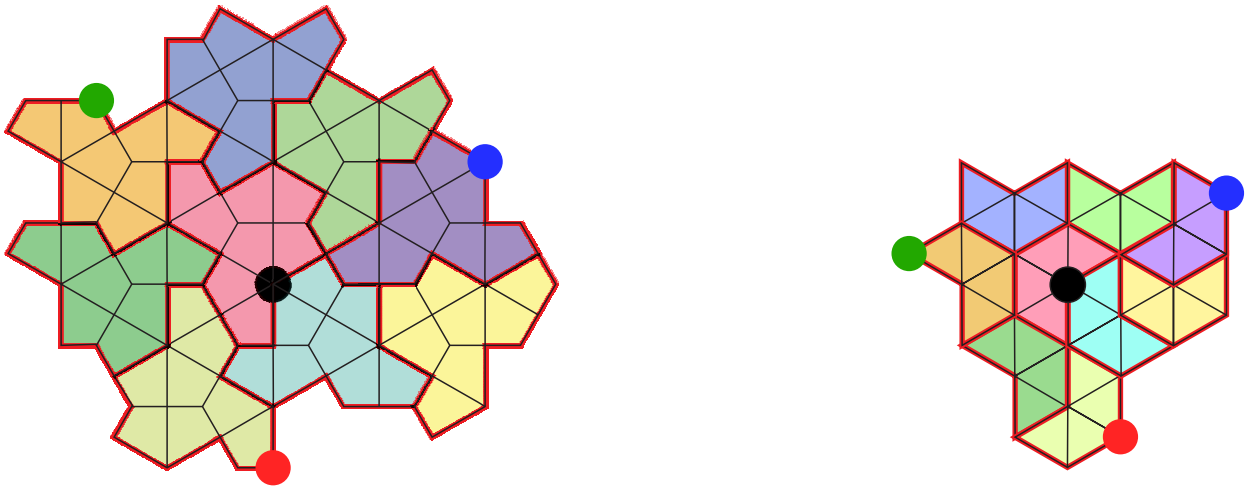
$$\sum_{i \in L} \vec{a}_i = \sum_{i \in S} \vec{a}_i = \vec{0}. \tag{2}$$

To see this, it is enough to focus on simple cycles since every cycle breaks into a union of simple cycles. Each simple cycle is the boundary of a topological disk that is a union of the tiles of a finite subset $\mathcal{F} \subseteq \mathcal{T}$, and without loss of generality we may assume that the cycle is oriented clockwise. We use induction on the number $k = |\mathcal{F}|$ of enclosed tiles. The claim is trivial if $k = 0$. Assume then $k > 1$, and let $t \in \mathcal{F}$ be an enclosed tile whose edge is on the path. Any edge of t that is tracked by the path is tracked in the clockwise orientation. Let $\vec{b}_1, \ldots, \vec{b}_{14}$ be the edge vectors of t in the opposite, counter clockwise direction. By (H1), both the long vectors and the short vectors among $\vec{b}_j$ add up to zero. We merge the collections of vectors $\vec{a}_i$ and $\vec{b}_j$. For each edge of t that is on the path there is the opposite vector among $\vec{b}_j$. Cancelling such pairs, the remaining vectors can be grouped into simple cycles that enclose subsets of $\mathcal{F} \setminus \{t\}$. By the inductive hypotheses the sums of the long and the short vectors among them both sum up to zero. So overall the sums of the long and the short vectors among $\vec{a}_i$ and $\vec{b}_j$ add up to zero, which implies the claimed equality (2).

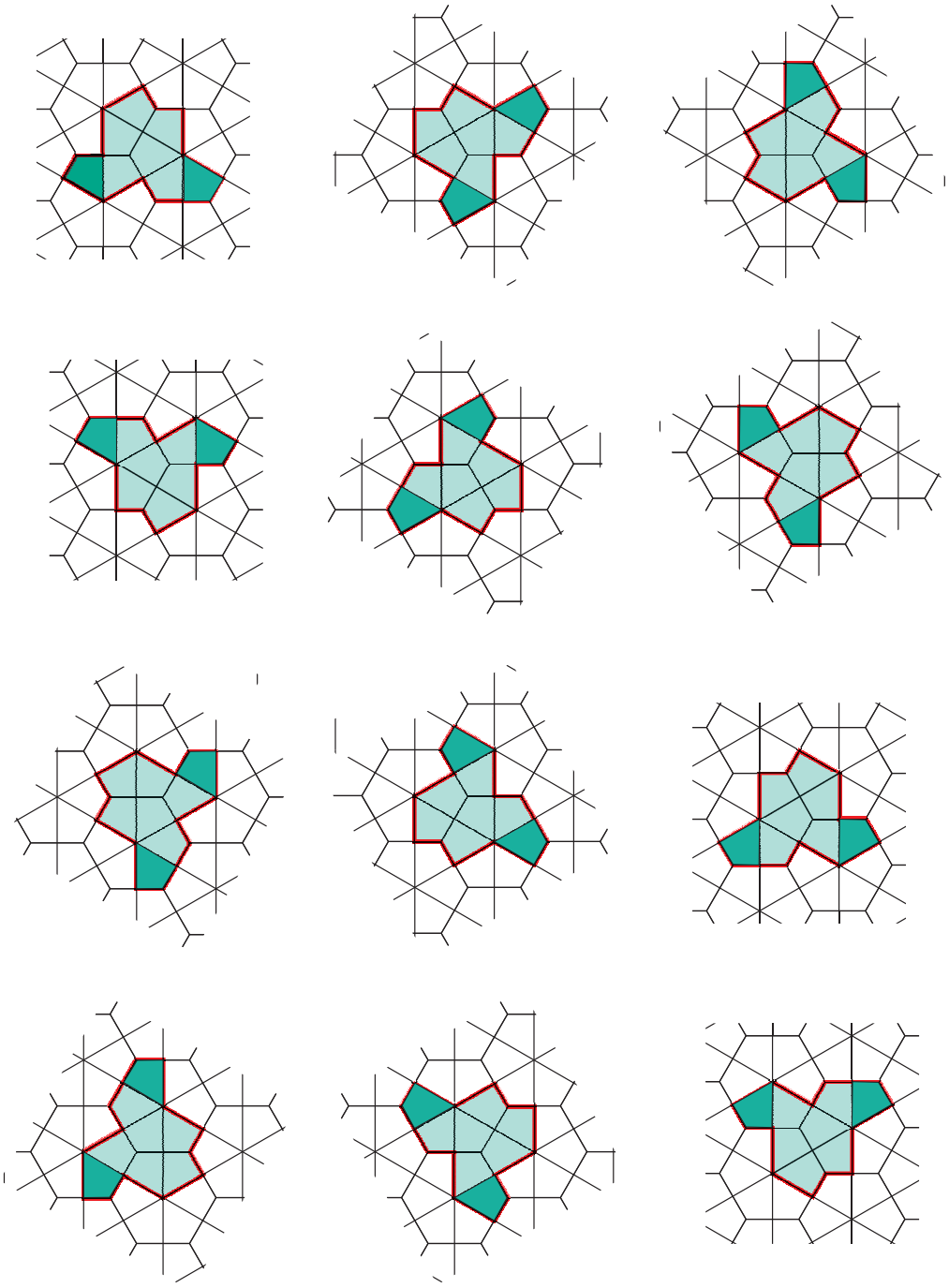
It now follows that scaling the long and the short edges of the hat by factors $a \geq 0$ and $b \geq 0$, respectively, transforms the tiling $\mathcal{T}$ into a tiling $\mathcal{T}'$ by the transformed tiles: Any vertex $\vec{v} \in \mathcal{V}$ is reached from $\vec{0}$ by a path $p = \vec{v}_0, \vec{v}_1, \ldots, \vec{v}_n$ in $\mathcal{T}$ where $\vec{v}_0 = \vec{0}$ and $\vec{v}_n = \vec{v}$. The corresponding vertex $\vec{v}'$

in $\mathcal{T}'$ is reached from $\vec{0}$ by the path p' obtained from p by scaling the long and short edges of the path by factors a and b, respectively. This results in the same $\vec{v'}$ regardless of the choice of p: If q is another path from $\vec{0}$ to $\vec{v}$ then the path p followed by q reversed is a cycle, and as proved above, the long and the short edges of the cycle sum up to zero. Thus the scaled long and short edges also sum up to zero, and therefore the paths p' and q' obtained by scaling p and q lead from $\vec{0}$ to the same point $\vec{v'}$. Thus the deformed tiling is well-defined.

The following picture illustrates a piece of a hat tiling and the corresponding tiling by the chevrons. The figure also shows four sample vertices of the hat tiling and the corresponding vertices in the chevron tiling.



Skipping the detailed proof, we note that all valid hat tilings of $\mathbb{R}^2$ must be such that the copies of the hat are aligned on the grid of kites. So in the following we only consider tilings where the hats are aligned on a given fixed grid of kites. This means that there are 12 possible orientations of the hat: six obtained by rotating the hat, and another six obtained by rotating the reflected hat:

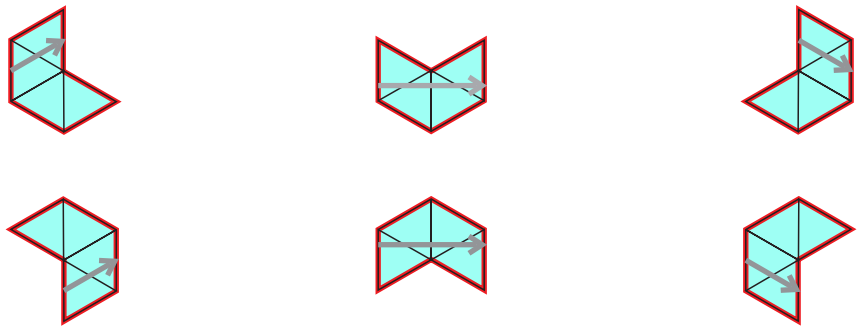


A notable property of the hat tile is that while the kites in the hat come in all 6 available orientations, one opposite pair of kite orientations is used twice. (In the illustration above these excess kites are indicated in darker color.) Contrasting this imbalance with the fact that the grid itself has kites in all six orientations in equal proportions, we see that in valid tilings there are three groups of four orientations of the hat (the columns of the illustration above) such that

- (H3) in a valid tiling, one third of the hats come oriented as in each of the three column in the picture above.

Recall that the long edges of hats are along i-lines for $i \in \{1, 2, 3\}$. Observe that the pairs of parallel long edges of the hat are always on consecutive i-lines. These lines are are at distance 3 from each other. Denote $D = \frac{3}{2}$ so that the i-lines repeat with distance $2D$.

Deforming the hat into a chevron preserves the orientations of the long edges. From the 12 possible orientations of the hat we obtain only six different orientation of the chevron, due to the reflection symmetry of the chevron:



The three columns in this picture correspond to the three columns in the picture of hat orientations before. It follows that

- (H4) a chevron tiling that is obtained by deforming a hat tiling has one third of its chevrons from each of the three columns above.

The chevrons are also aligned with a grid of equilateral triangles, formed by three families of parallel lines in the same directions as the i-lines in the grid of kites. We also call these i-lines, for $i \in \{1, 2, 3\}$. The sides of the triangles are of length $\sqrt{3}$, and so the consecutive i-lines of chevrons are at distance $\frac{3}{2}$ from each other,*i.e.*, the distance is D, half of the distance of i-lines of hats.

Consider, for example, the pair of vertical ($i = 1$) edges of a chevron: if oriented as in the first column above, the edges are on consecutive 1-lines but the rightmost edge is higher (by $\sqrt{3}/2$) than the leftmost edge. (See the grey vector in the picture above.) In the third column the rightmost edge is lower by the same amount. In the middle column the vertical edges are on the same height but they are not on consecutive 1-lines, but at double distance $2D$. We have that, on the average, the vertical position of the two vertical edges is the same, and the average horizontal distance between them is $\frac{4}{3}D$. Due to symmetry, edges in the other two directions behave similarly.

In the following we prove that the hat tile does not admit a two-periodic tiling. Assume the contrary: suppose $\mathcal{T}$ is a two-periodic tiling using the hat. Let $\vec{p}$ and $\vec{q}$ be generators of the periods of $\mathcal{T}$, meaning that $i\vec{p} + j\vec{q}$ are precisely its vectors of periodicity, for $i, j \in \mathbb{Z}$. Denote $\mathcal{P} = \mathbb{Z}\vec{p} + \mathbb{Z}\vec{q}$ for the set of the periods of $\mathcal{T}$. Let $\mathcal{V} \subseteq \mathbb{R}^2$ be the set of vertices of $\mathcal{T}$, and assume that $\vec{0} \in \mathcal{V}$. Note that

$$\vec{v} \in \mathcal{V} \iff \vec{v} + \vec{p} \in \mathcal{V} \iff \vec{v} + \vec{q} \in \mathcal{V},$$

which, together with $\vec{0} \in \mathcal{V}$, implies that $\mathcal{P} \subseteq \mathcal{V}$.

We deform the hat and the periodic tiling $\mathcal{T}$ by scaling factors $a = 1$ and $b = 0$, as discussed above. Let $\mathcal{T}'$ be the deformed tiling (by chevrons) and let $\mathcal{V}'$ be its vertex set. Let $f : \mathcal{V} \longrightarrow \mathcal{V}'$ give for each vertex the corresponding vertex in the deformed tiling, *i.e.*, if p is a path in $\mathcal{T}$ from $\vec{0}$ to $\vec{v} \in \mathcal{V}$ then the corresponding deformed path p' leads in $\mathcal{T}'$ from $\vec{0}$ to $f(\vec{v})$. (Note that f is not one-to-one: removing short edges merges the vertices they connect.)

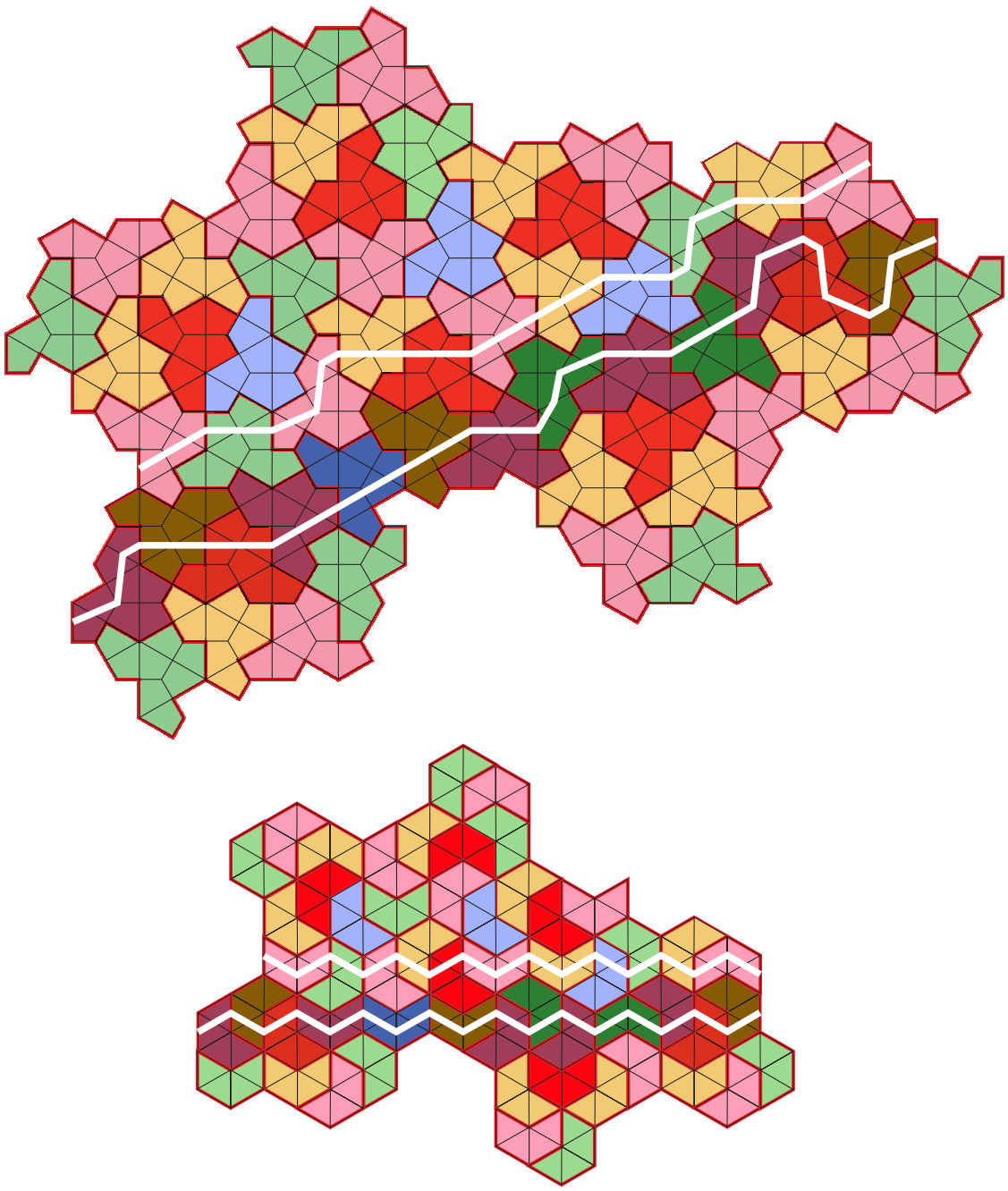
For any $\vec{x} \in \mathcal{V}$, choose a path $p_{\vec{x}}$ in $\mathcal{T}$ from $\vec{0}$ to $\vec{x}$ and let $p'_{\vec{x}}$ be the corresponding deformed path in $\mathcal{T}'$ from $\vec{0}$ to $f(\vec{x})$. Let $\vec{v} \in \mathcal{P} \subseteq \mathcal{V}$ be any period of $\mathcal{T}$, and let $\vec{x} \in \mathcal{V}$ be arbitrary. Since $\vec{v} \in \mathcal{P}$, there is a path q in $\mathcal{T}$ from $\vec{v}$ to $\vec{v} + \vec{x}$ that is a translation by $\vec{v}$ of the path $p_{\vec{x}}$. The path $p_{\vec{v}}$ followed by the path q is a path from $\vec{0}$ to $\vec{v} + \vec{x}$ whose deformation leads in $\mathcal{T}'$ from $\vec{0}$ to $f(\vec{v}) + f(\vec{x})$. We see that $f(\vec{v} + \vec{x}) = f(\vec{v}) + f(\vec{x})$. In particular, we have that $f(\vec{v})$ is a period of $\mathcal{T}'$ since for any vertex $f(\vec{x}) \in \mathcal{V}'$ also $f(\vec{v}) + f(\vec{x})$ is in $\mathcal{V}'$.

Moreover, it follows that $f(i\vec{p} + j\vec{q}) = if(\vec{p}) + jf(\vec{q})$ for all $i, j \in \mathbb{Z}$ where $\vec{p}, \vec{q}$ are the generating periods of $\mathcal{T}$. This means that f is linear among the periods. Let us denote by $\hat{f} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ the unique linear map that coincides with f on $\mathcal{P}$, that is, the unique linear function that maps $\vec{p} \mapsto f(\vec{p})$ and $\vec{q} \mapsto f(\vec{q})$. Note that f and $\hat{f}$ do not need to coincide on $\mathcal{V} \setminus \mathcal{P}$. The linear function $\hat{f}$ is one-to-one: it has full rank since otherwise $f(\mathcal{P})$ would be on a single line, which is clearly not the case. (If $f(\mathcal{P}) \subseteq \ell$ for some line ℓ, then for any two hats in $\mathcal{T}$ that are transitive under some translational symmetry of $\mathcal{T}$ the corresponding deformed tiles in $\mathcal{T}'$ are on a common line parallel to ℓ. There are only finitely many transitivity classes under translational symmetries in $\mathcal{T}$ but clearly there is an infinite set of chevrons in $\mathcal{T}'$ that are pairwise not on such a common line, a contradiction.)

Our next goal is to prove that the linear function $\hat{f}$ is a similarity, *i.e.*, distances of points are scaled by the same constant: we argue that there exists a constant c such that for any $\vec{x}, \vec{y} \in \mathbb{R}^2$ it holds that $d(\hat{f}(\vec{x}), \hat{f}(\vec{y})) = d(\vec{x}, \vec{y})/c$.

Consider any of the three directions $i \in \{1, 2, 3\}$. A *de Bruijn segment* on a hat tile in direction i is a line drawn inside the tile connecting the centers of the two long edges of the hat that are parallel to i-lines. Similarly we define de Bruijn segments of chevrons. There is one de Bruijn segment drawn in each tile in each direction i. Here are examples in the vertical direction $i = 1$:

In the valid tiling $\mathcal{T}$, the de Bruijn segments of two tiles sharing a long edge continue across that edge, thus defining infinite *de Bruijn lines* in directions i across the entire tiling. Each tile is crossed by a unique de Bruijn line in each direction i, and the lines in the same direction do not cross each other. Let us call the set of tiles on the same de Bruijn line in direction i an *i-strip*. Similarly we define de Bruijn lines and i-strips in tilings by chevrons. The following picture shows a patch of a tiling by hats, and two de Bruijn lines across the patch in the vertical direction $i = 1$. Tiles of the i-strip of the lower line are rendered dark. The lower patch is the same drawing in the corresponding tiling by chevrons.



Note that any translational symmetry τ of $\mathcal{T}$ must map i-strips onto i-strips, preserving the order of tiles on the strips: if $t_2 = \tau(t_1)$ then the tile following (preceding) t_1 on its i-strip will be mapped on the tile following (preceding, resp.) t_2 on its i-strip.

Fix direction i and consider one i-strip. The strip contains infinitely many tiles. Because $\mathcal{T}$ is two-periodic, there are only finitely many tiles that are not transitive under translational symmetries of $\mathcal{T}$. Thus the strip contains distinct tiles t_1, t_2 such that $t_2 = \tau(t_1)$ for some translational symmetry τ of $\mathcal{T}$. As noted above, this means that the strip is mapped by τ onto itself. As τ cannot change the order of i-strips, it follows that every i-strip is mapped onto itself by τ.

A similar reasoning works with all directions $i \in \{1, 2, 3\}$. For each i, let $\vec{u_i}$ be a unit vector parallel to i-lines, chosen so that $\vec{u_1}$, $\vec{u_2}$ and $\vec{u_3}$ are at 120° angles to each other. For each i, let $\vec{v_i}$ be a vector of length $2D$, perpendicular to i-lines, again chosen so that $\vec{v_1}$, $\vec{v_2}$ and $\vec{v_3}$ are at 120° angles to each other, *i.e.*, so that $\vec{v_1} + \vec{v_2} + \vec{v_3} = \vec{0}$.

As seen above, for each $i \in \{1,2,3\}$ there is a translational symmetry of $\mathcal{T}$ that maps i-strips onto themselves. Let such a translation in direction $i \in \{1,2,3\}$ be by the vector $\vec{p_i} = x_i\vec{v_i} + y_i\vec{u_i}$. Here x_i is a positive integer indicating how many tiles the translation moves forward on an i-strip. As multiples of periodicity vectors are also periodicity vectors, we can choose the periods $\vec{p_i}$ so that $x_1 = x_2 = x_3$ is the same positive integer k.

Consider then the i-strips on the corresponding chevron tiling $\mathcal{T}'$. These are precisely the i-strips of $\mathcal{T}$ after the deformation. The translation by vector $f(\vec{p_i}) = \hat{f}(\vec{p_i})$ is a symmetry of $\mathcal{T}'$ that maps each i-strip onto itself: it shifts by k the tiles within each i-strip. Averaging over all finitely many i-strips with distinct translational transitivity classes we see – based on observation (H4) – that the vector $\hat{f}(\vec{p_i})$ must be perpendicular to the i-lines, and its length is $k \cdot \frac{4}{3}D$. The length is independent of i, and the directions for $i \in \{1,2,3\}$ are at 120° angles to each other, implying that $\hat{f}(\vec{p_1}) + \hat{f}(\vec{p_2}) + \hat{f}(\vec{p_3}) = \vec{0}$. Linearity of $\hat{f}$ thus means that $\hat{f}(\vec{p_1} + \vec{p_2} + \vec{p_3}) = \vec{0}$, and by the injectivity of $\hat{f}$ then $\vec{p_1} + \vec{p_2} + \vec{p_3} = 0$.

On the other hand, as $\vec{p_i} = k\vec{v_i} + y_i\vec{u_i}$ and $\vec{v_1} + \vec{v_2} + \vec{v_3} = \vec{0}$, we have that $y_1\vec{u_1} + y_2\vec{u_2} + y_3\vec{u_3} = \vec{0}$. Vectors $\vec{u_i}$ are unit vectors at angles 120° to each other, so that we must have $y_1 = y_2 = y_3$. In conclusion, vectors $\vec{p_i}$ have equal lengths and they are at 120° angles to each other, and their images $\hat{f}(\vec{p_i})$ under $\hat{f}$ have also equal lengths and they are at 120° angles to each other. This implies that $\hat{f}$ is a similarity map, i.e., an isometry followed by scaling by some constant $1/c$.

Using the fact (H2) that the area of the hat tile is 8/3 times the area of the chevron, we can even conclude the value of the similarity factor to be $c = \sqrt{8/3}$. Indeed, if there are m translational transitivity classes of hats in tiling $\mathcal{T}$, then the area of the parallelogram with sides $\vec{p}$ and $\vec{q}$ is m times the area of the hat tile. Based on $\mathcal{T}'$ then the area of the parallelogram with sides $\hat{f}(\vec{p})$ and $\hat{f}(\vec{q})$ is m times the area of the chevron tile. As $\hat{f}$ is a similarity with scaling c, the ratio of the areas of the parallelograms is c^2, while the ratio of the area of a hat to the area of a chevron is 8/3. This gives $c^2 = 8/3$.

Finally we show that scaling by $c = \sqrt{8/3}$ is not possible. Note that the periodicity vector $\vec{p}$ of $\mathcal{T}$ is a vector between two vertices of the underlying triangle grid formed by i-lines (as the translational symmetry is also a translational symmetry of the underlying grid of kites, so it maps sharp ends of kites to sharp ends – and the sharp ends are located at the vertices of the triangle grid.) On the other hand, the corresponding vector $\vec{p'} = \hat{f}(\vec{p})$ in $\mathcal{T}'$ is a vector between two vertices of the triangular grid of the chevron tiling. The triangles in the first grid have sides twice as long as in the second grid, so they can be subdivided into smaller triangles to get the second grid. Both vectors $\vec{p}$ and $\vec{p'}$ are thus connecting vertices of the same grid of small triangles. Due to the similarity scaling, the lengths of the vectors are related by $|\vec{p}|/|\vec{p'}| = \sqrt{8/3}$.

Let us prove that the length ratio $\sqrt{8/3}$ is not possible between two non-zero vectors connecting vertices of the same triangular grid. For the simplicity of notations, consider the grid of equilateral triangles containing vertices $(0,0)$ and $(1,0)$. The vertices of the grid have coordinates $i(1,0) + j(1/2, \sqrt{3}/2)$ for $i, j \in \mathbb{Z}$. If d is the distance of such a vertex from $(0,0)$ then $d^2 = (i + j/2)^2 + (j\sqrt{3}/2)^2 = i^2 + j^2 + ij$. Thus the squares of the lengths of vectors between vertices come from the set $S = \{i^2 + j^2 + ij \mid i, j \in \mathbb{Z}\}$. Elements of S are integers. Let us show that for any non-zero $s \in S$ one has $8/3s \notin S$. If i is odd or j is odd (or both are odd) then $s = i^2 + j^2 + ij$ is odd. If i and j are both even then dividing both by 2 gives that $s/4 \in S$. This means that any even $s \in S$ is divisible by 4, and then $s/4 \in S$. Repeatedly dividing any even $s \in S$ by 4 one hence eventually reaches an odd number. In conclusion: The largest power of 2 that divides any non-zero $s \in S$ is even, and therefore $s \in S \Longrightarrow 8/3s \notin S$.

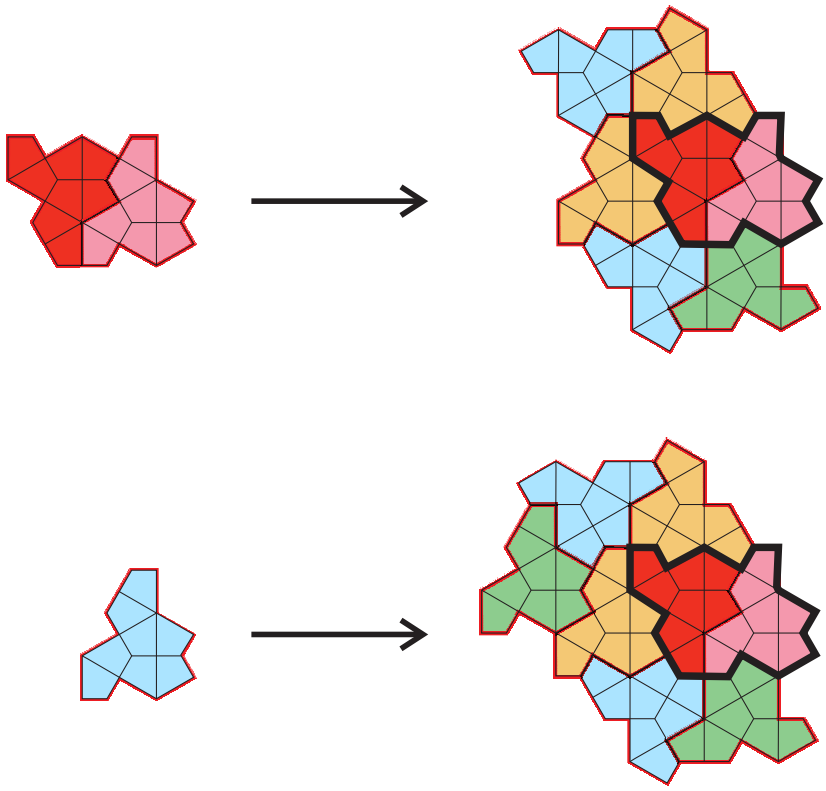
This shows that $|\vec{p}|^2 = 8/3|\vec{p'}|^2$ is not possible, and we have the following:

Theorem 7.6 *The hat tile does not admit any two-periodic tiling.*

By Theorem 7.5 (the periodicity theorem), the hat tile does not even admit a one-way periodic tiling.

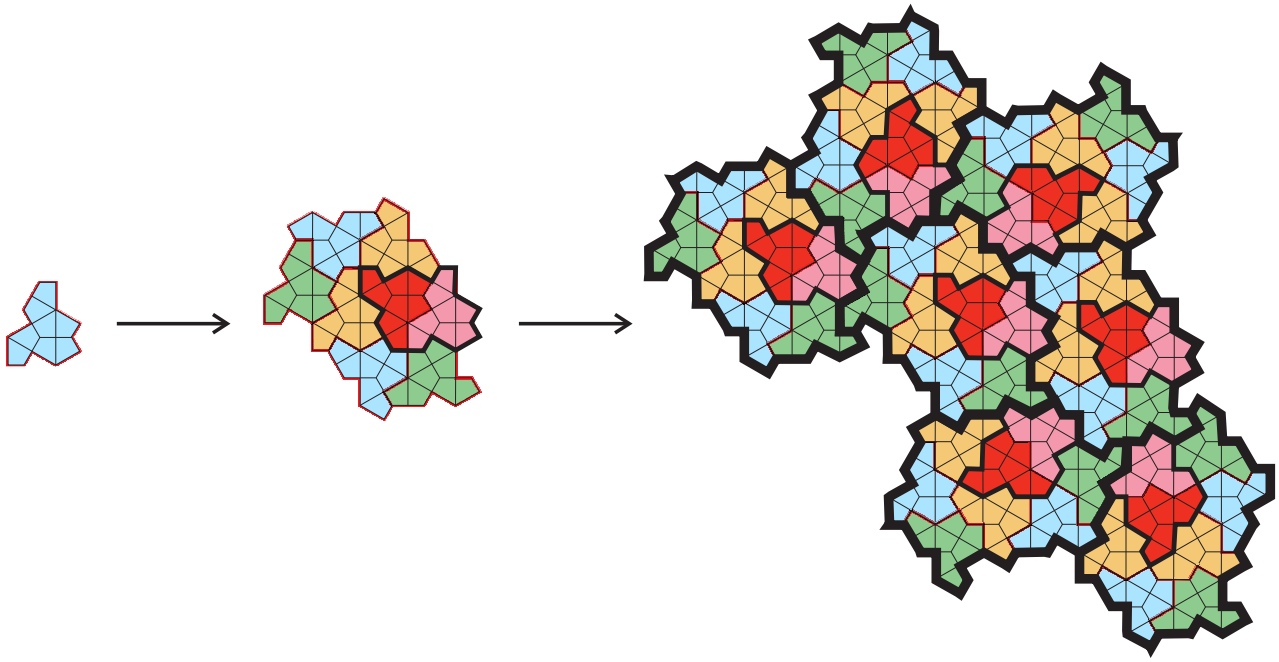
We skip the details of the proof that the hat tile admits a tiling. One can generate a tiling as follows. Define first patch P of two tiles by attaching the hat tile and the flipped over variant of the hat (upper left

patch on the figure below). Let Q be a patch consisting of a single hat tile (lower left patch). Then define a substitution on patches P and Q that replaces each P and Q by super-P and super-Q, respectively (upper right and lower right in the figure):



Both supertiles contain one copy of P, shown in red/pink in the figure, and several patches Q. None of the patches are flipped over, so the red tile inherited from Q is the only tile of the super patches that is a flipped over variant of the hat.

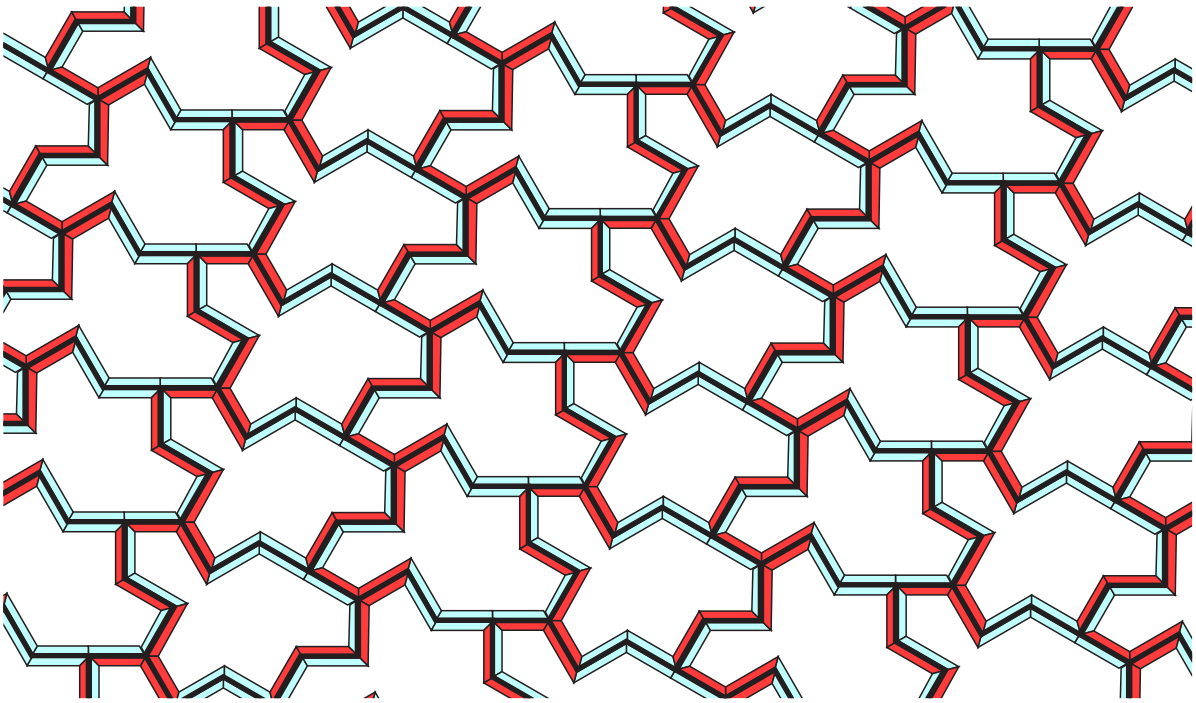
This substitution can be iterated, thus obtaining as a limit a tiling of the plane by hats. The following picture shows the creation of the second level super-P patch:



Note that the tiles in dark red are the only flipped over hat tiles. The fact that the hat tile does not admit a periodic tiling (Theorem 7.6) can also be proved analogously to our previous substitution examples: any valid tiling can be uniquely decomposed into super-patches. There are several cases to consider and we skip the proof.

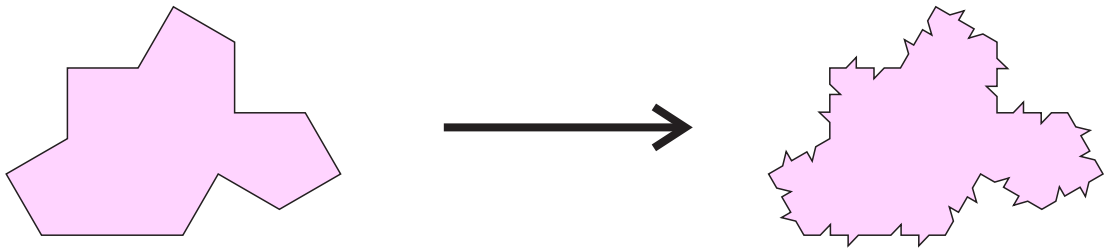
Some final remarks: Deforming the hat with scaling factors $a \geq 0$ and $b \geq 0$ on long and short edges produces "equivalent" tiles that are also aperiodic, except when (i) $a = 0$, or (ii) $b = 0$ or (iii) $b/a = \sqrt{3}$. In cases (i) and (ii) the long and short edges vanish, respectively, and in case (iii) the long and short edges become equally long. In all cases deforming a tiling by the hat becomes a tiling by the deformed tile, so all deformed tiles admit monohedral tilings. Conversely, in all cases except (i), (ii) and (iii) a valid tiling by the deformed tiles is necessarily edge-to-edge in such a manner that long edges meet long edges and short edges meet short edges. In this situation the inverse deformation can be done back from the deformed tile to the hat. If the deformed tile would admit a periodic tiling, the inverse deformation of this tiling would produce a periodic tiling by the hat, which by Theorem 7.6 does not exist. So the deformed tile does not admit a periodic tiling.

Note that in cases (i), (ii) and (iii) the inverse deformation is not well defined on tilings. In particular, in case (iii) the tile admits a valid tiling where two neighboring tiles meet with their "long" and "short" edges against each other. In the periodic tiling below the "long" edges are shown red, short edges are shown as light blue.



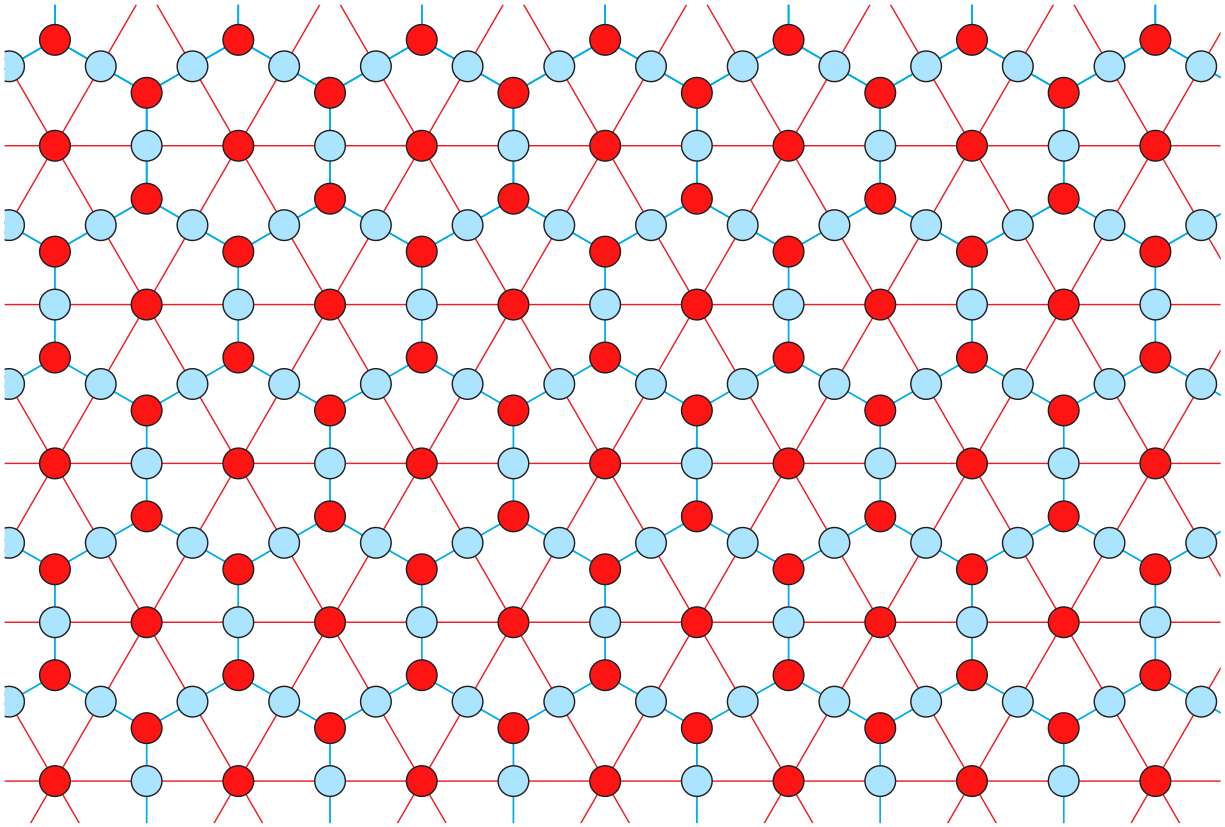
Note that there are several instances where a red edge meets a blue edge. In this situation the inverse deformation breaks the tiling. So this periodic tiling, of course, does not provide a periodic tiling by the hat.

Note that in the periodic tiling above both even and odd (=flipped over) variants of the tile are used. It turns out (proof skipped) that the tile also admits a tiling using only the even variants, and that all tilings using only the even variants are non-periodic! Using a standard bumps and dents construction on the edges we can prevent the even and the odd variants from appearing in the same tiling. In this construction the new shape of the edge is symmetric under a half turn so that the even tiles still match with each other along their edges:

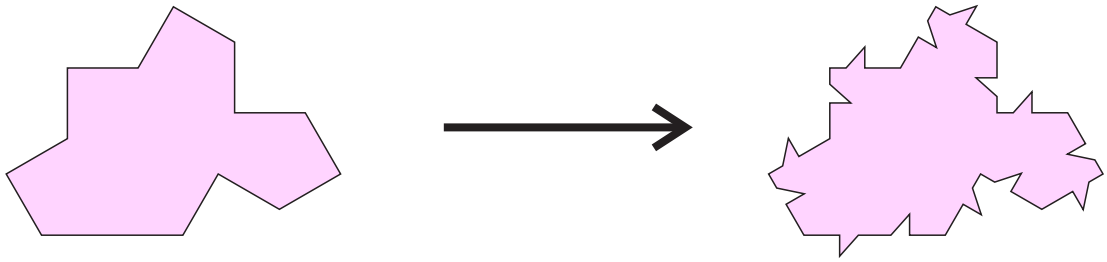


This tile is thus aperiodic monotile in the following strong sense: it admits a tiling, no tiling involves both even and odd variants of the tile, and there is no valid periodic tiling. The tile is called *spectre* (as is any of the analogous variants where differently curved edges are used to prevent even and odd tiles next to each other).

One more interesting observation can be made: The original kite grid is bipartite in the sense that we can color red and color blue those vertices where edges meet at angles that are multiples of 120° and multiples of 90°, respectively, and each edge then connects a red and a blue vertex:

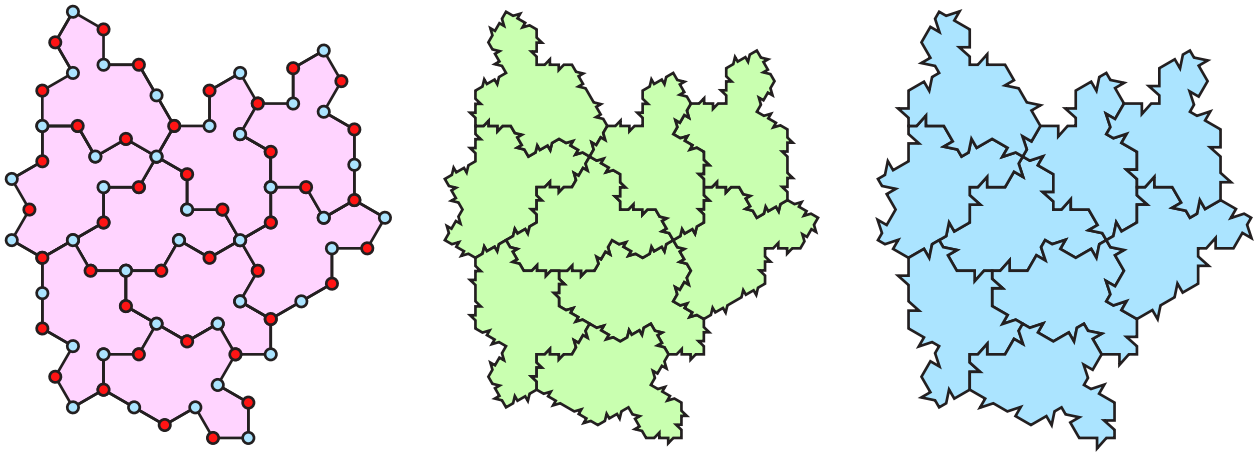


Thus, if we color vertices of the hat tile alternatingly red and blue, in any valid tiling by the hat the meeting vertices have the same color. The same is then true on the deformed hat. Consequently, the bumps and dents construction above can be simplified so that the new shape of the edge does not need to be symmetric under a half turn, but instead the edges of the tile are alternatingly the shape and its half turned variant:



This tile has the same spectre properties: it admits a tiling, no tiling involves both even and odd variants of the tile, and there is no valid periodic tiling.

The following picture shows corresponding patches using the deformed hat (left, including coloring of the vertices), spectre with half turn symmetric edge shapes (middle) and spectre with the simpler edge shapes that are not symmetric but alternate in consecutive edges (right).



7.5 Open problems

The aperiodic monotile "hat" discussed above is a 13-gon. A natural follow-up problem is to try to reduce the number of sides. Quadrilaterals tile the plane periodically, so the smallest possible number of sides on a polygonal aperiodic monotile is at least 5.

Open problem *What is the smallest n such that there exists an aperiodic n-gon ? Does there exist an aperiodic pentagon ?*

It is known that an aperiodic n-gon cannot be *convex*: any convex polygon that admits a tiling also admits a periodic tiling. This is clear for $n \leq 4$ as all triangles and quadrilaterals tile the plane periodically. Using Euler's formula for planar graphs ($v-e+f=2$ where v, e and f stand for the number of vertices, edges and faces of the graph, respectively) it is fairly easy to see that no convex n-gon with $n \geq 7$ admit any tiling. Convex hexagons ($n=6$) were analysed by K.Reinhardt in 1918: there are only three "types" of convex hexagons that tile the plane, and they all tile periodically. Convex pentagons are harder to analyse. There are 15 types that tile – the 15th type was discovered using a computer search as late as in 2015. In 2017 M.Rao gave a computer-assisted proof that there are no more than 15 types of convex pentagons that tile the plane. They all tile periodically.

The construction from Wang tiles to polygons with bumps and dents is clearly effective, which means that it can be executed mechanically by an algorithm. Consequently, the undecidability results proved for Wang tiles hold for polygonal prototiles as well. In decision problems whose input consist of polygons we must use a finite way of representing the polygons. A natural way is to assume that the input polygons are such that the vertices of the polygons have rational coordinates, and the encoding then is a list of

consecutive vertices. The bumps and dents can be done with rational coordinates so we immediately get the following undecidability results, corresponding to Theorems 5.7 and 5.8.

Theorem 7.7 *The following decision problems are undecidable:*

- *"Does a given protoset of polygons with rational coordinates admit a periodic tiling ?",*
- *"Does a given protoset of polygons with rational coordinates admit a tiling ?".*

Proof. Suppose an algorithm A exists for one of the given problems. Then we can solve the analogous problem on Wang tiles: For the given Wang protoset we use the bump/dent construction to form a protoset of polygons, and we give this protoset as input to the hypothetical algorithm A. The answer from algorithm A tells whether the Wang protoset admits a (periodic) tiling. This contradicts Theorem 5.7 or 5.8.

□

It is not known if there exists a decision algorithm to determine if a given single polygonal prototile admits a valid (periodic) tiling. Note that with Wang tiles the question is trivial, as for every k there are only a finite number of non-isomorphic protosets with k tiles, and consequently the decision problems are decidable. But there are infinitely many different polygons, so there is no trivial reason why there would exist an algorithm to tell even if a single tile admits a tiling. The aperiodic monotile "hat" provides the necessary pre-condition for undecidability.

Open problem *Is the following decision problem decidable ?*

- *"Does given single polygon with rational coordinates admit a tiling ?"*

What about the same question restricted to pentagons ?

□

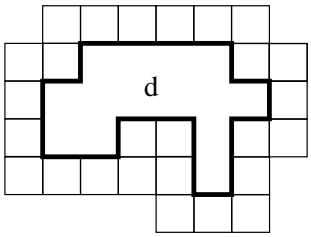
It has been shown by N. Ollinger that the tiling problem is undecidable among protosets that contain 5 polyominoes. (A polyominoe is a tile obtained by edge-to-edge attachments of any number of unit squares to each other.) It has also been recently established that the problem remains undecidable for five polyominoes even if only translations are allowed, that is, the tiles must be placed in the given orientation (Y.Kim). On the other hand, the tiling problem is known to be decidable for single polyominoes if only translations are allowed (Wijshoff, van Leeuwen).

In particular, Ollinger's result implies the undecidability of the tiling problem among sets of 5 polygons. This was recently improved to 3 polygons by Demaine and Langerman.

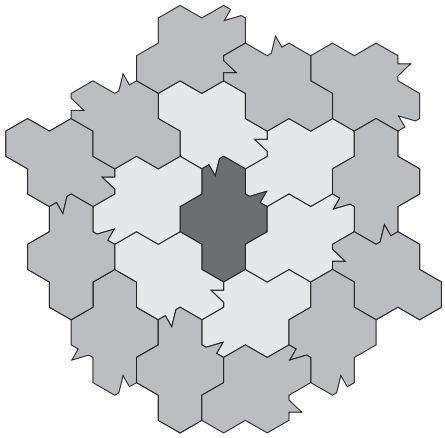
As a related question, consider the following problem by H. Heesch. Given a prototile t that does not admit a tiling of the plane, the Heesch number of t is the maximum number of times the tile can be completely surrounded by copies of t. More precisely, for a topological disk $d \subseteq \mathbb{R}^2$, a *corona* of d is a collection C of tiles, all congruent to t, such that

1. the interiors of the elements of C are pairwise disjoint, and disjoint from d, and
2. $d \cup \bigcup_{s \in C} s$ is a topological disk whose interior contains d.

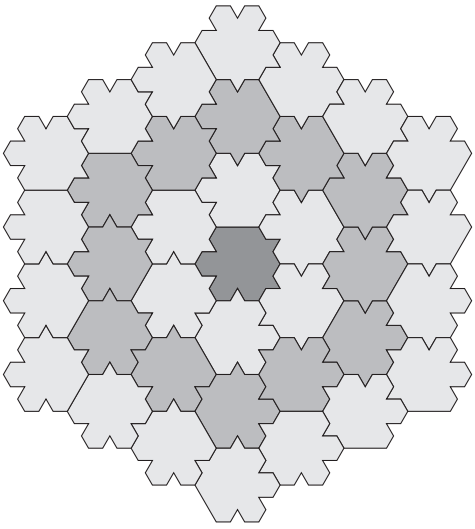
In other words, tiles in the corona C surround set d completely. For example, the squares in the following figure form a corona of the set d in the middle:



We can define a second corona of d as a corona of the set that is the union of d and its first corona. Inductively, a $k+1$'st corona is a corona of the topological disk formed by d and its first k coronas. In the Heesch problem we start with a single copy of t and form its 1st, 2nd, 3rd, etc. coronas. If the k'th corona exists for every k then by Theorem 7.4 the entire plane can be tiled. But if t does not admit a plane tiling then there exists the largest k such that the first k coronas exist. This k is called the *Heesch number* of tile t. The following figure illustrates two coronas of a tile:

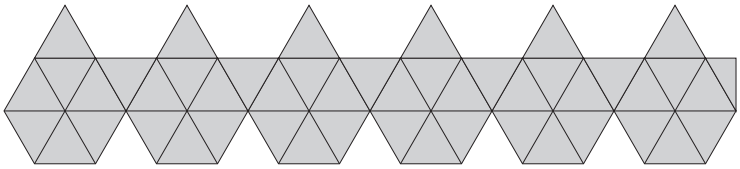


A regular hexagon with incoming arrows on three sides and outgoing arrows on two sides admits three coronas:

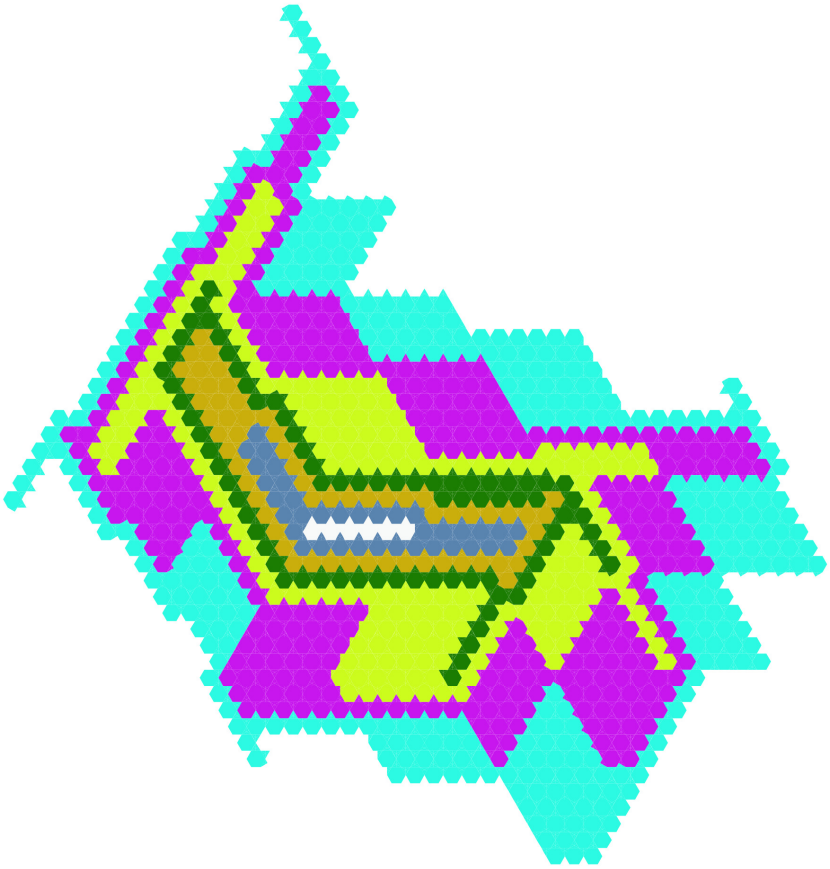


(In the picture, the arrows are represented by bumps and dents.) Due to the imbalance in the number of incoming and outgoing arrows, it is easy to apply the argument used earlier in the homeworks to conclude that the full plane cannot be tiled by this tile.

At the present time the largest known Heesch number is six. This is is a tile reaching that number:



Here's a picture showing the maximum number of coronas :



[Bašić, Bojan – Smaller version of https://en.wikipedia.org/wiki/Heesch]

Open problem *Does there exist number k such that the Heesch number of every tile that does not admit a tiling is at most k ? If such a k exists, what is the smallest such k ?*

□

Note that if the Heesch numbers are bounded by some constant k then there is an algorithm (at least in any reasonable set-up such as edge-to-edge tilings by polygons where one can try all possible coronas) to determine if a given single tile admits a tiling: To test if a tiling exists, all we need to do is to try all possible ways of building $k+1$ coronas. A valid tiling exists if and only if $k+1$ coronas exist.