Groups, Tilings and Finite State Automata

Summer 1989 AMS Colloquium Lectures

William P. Thurston

Research Report GCG 1

Groups, tilings, and finite state automata
Summer 1989 AMS Colloquium lectures

(Version 1.5, July 20, 1989)

WILLIAM P. THURSTON

§1. INTRODUCTION

These four lectures will develop some ideas involving the geometry of groups, tilings (primarily of the plane), finite state automata, and dynamical systems. They are grouped into three related subjects which are tied together by common themes, but are sufficiently independent that it should be possible to understand them independently.

The subject of the first lecture is a connection between tilings of the plane and the geometry of groups discovered by Conway a number of years ago, but only recently been discussed in print ([**Th0**], [**Conway Lagarias**]). It develops a necessary condition for a region in the plane to be tiled by a given collection of tiles, in terms of combinatorial group theory.

The second lecture also concerns tilings, but from a different point of view: the subject is the theory of self-similar tilings of the plane and of other Euclidean spaces. Many examples and constructions will be discussed. The main result is a characterization of the complex expansion constants for selfsimilar tilings. This subject is closely related to the theory of Markov partitions for dynamical systems and finite state automata. In a certain sense it may be thought of as a complexification of the Perron-Frobenius theorem and its 'converse' of D. Lind.

Word processing on groups, or the theory of automatic groups, is the subject of the last two lectures. This theory has been developed over the last few years primarily in joint work of Jim Cannon, David Epstein, Derek Holt, Mike Paterson, and me ([**CEHPT**]). An automatic groups admits an algorithm of a rather simple type which will tell when two words in generators for the group represent the same element of the group (*i.e.*, an algorithm for the word problem of the group.) Moreover, the algorithm is so special and so simple that questions about the algorithm can be algorithmically handled: in particular, there is an algorithm which, given a presentation for an automatic group, will construct an algorithm as above for the word problem.

Automatic groups are closely tied to the theory of finite state automata, and the investigation of them is partly motivated by the successful applications which finite state automata have found to practical and theoretical problems in computer science, combined with the need to be able to handle algorithmically actual finitely-presented infinite groups (in particular, fundamental groups of 3-manifolds.) Many word-processors — for example the unix utilities grep, egrep, sed, vi, *etc.* — construct a finite-state automaton when you ask it to search for a certain pattern, and many compilers directly use the theory of finite state automata at early stages of their tasks (lexical and syntactical analysis). Besides theoretically analyzing the issues involved in automatic groups, we have been developing computer programs to carry out 'word-processing on groups'. Automatic groups

are more general than hyperbolic groups in the sense of Gromov. At least most of the small-cancellation groups are automatic.

An automatic structure for a group in general produces a kind of self-similar tiling of a certain 'sphere at infinity' for the group; in particular examples, this space is actually a 2-sphere.

These notes are preliminary. Although some portions have been written carefully and in fair detail, there are other portions are sketchy and hastily written, and some topics have been left out altogether.

The next portion of this text, concerning Conway's tiling groups (§2 — §7) is substantially a reprint from an article to appear in the January 1990 issue of the *American Mathematical Monthly*, [**Th0**]. This will be a special issue on geometry.

§2. Conway's tiling groups

The problem of deciding whether a given finite set of tiles will tile the plane is an undecidable question — that is, there is no general well-defined procedure which will answer the question. The same question for a finite region in the plane, when appropriately formulated, is decidable, but it is not easy: it is what computer scientists call an NP-complete question. In practice, it is often hard to do.

John Conway discovered a technique using infinite, finitely presented groups that in a number of interesting cases resolves the question of whether a region in the plane can be tessellated by given tiles. The idea is that the tiles can be interpreted as describing relators in a group, in such a way that the plane region can be tiled, only if the group element which describes the boundary of the region is the trivial element 1.

Of course, the word problem for a finitely-presented group (the problem of deciding whether or not two given words represent identical elements in the group) is also an undecidable question. The ability to answer the tiling questions depends in part on the ability to understand particular group presentations . . . .

§3. Group graphs

A convenient way to describe the construction is by means of the *Cayley graph* or *graph* of a group. If G is a group, then its graph $\Gamma(G)$ with respect to generators $g_1, g_2 \ldots, g_n$ is a directed graph whose vertices are the elements of the group. For each vertex $v \in \Gamma(G)$, there will be n outgoing edges, labeled by the generators, and n incoming edges: the edge labeled g_i connects v to vg_i.

As a first example, the graph of $\mathbf{Z}^2$ with respect to standard generators $\langle x, y | xyx^{-1}y^{-1} \rangle$ is the standard grid in the plane (as in graph paper).

The graph of a group is an answer to the question, 'what does a group look like?' which generally is carefully avoided in introductory courses. Note however that the graph of a group depends on the choice of generators, and the appearance can change considerably with a change of generators: the group graph tells what a group with a little extra structure looks like.

It is convenient to make a slight modification of this picture when a generator g_i has order 2. In that case, instead of drawing an arrow from v to vg_i and another arrow from

vg_i back to v, we draw a single undirected edge labeled g_i. Thus, in a drawing of the graph of a group, if there are undirected edges, it is understood that the corresponding generator has order 2.

The graph of a group is automatically homogeneous: for every element $g \in G$, the transformation $v \to gv$ is an automorphism of the graph. Every automorphism of the labeled graph has this form. This property characterizes graphs of groups: a graph whose edges are labeled by a finite set F such that there is exactly one incoming and one outgoing edge with each label at each vertex is the graph of a group if and only if it admits an automorphism taking any vertex to any other.

Whenever R is a relator for the group, that is, a word in the generators which represents 1, then if you start from $v \in \Gamma$ and trace out R, you get back to v again. If G has presentation

$$G = \langle g_1, g_2, \ldots, g_n | R_1 = 1, R_2 = 1, \ldots, R_k = 1 \rangle$$

the graph $\Gamma(G)$ extends to a 2-complex $\Gamma^2(G)$: sew k disks at each vertex $v \in \Gamma(g)$, one for each relator R_i, so that its boundary traces out the word R_i. An exception is made here for relations of the form $g_i^2 = 1$, since this relation is already incorporated by drawing g_i as an undirected edge. The 2-complex $\Gamma^2(G)$ is simply-connected: that is, every loop in $\Gamma^2(G)$ can be contracted to a point. In fact, if the loop is an edge path, the sequence of edges it follows describes a word in the generators. The fact that the path returns to its starting point means that the word represents the identity. A proof that this word represents the identity by making substitutions using the relations R_i can be translated geometrically into a homotopy of the path in $\Gamma^2(G)$.

As a very simple example, the symmetric group S_3 is generated by the transpositions $a = (12)$ and $b = (23)$. They satisfy the relation $(ab)^3 = 1$. The graph is a hexagon, with undirected edges, alternately labeled a and b.

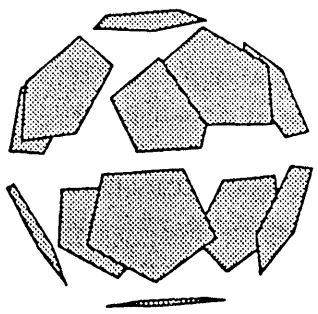

Figure 3.1. Soccerball. *A soccerball is constructed from 12 pentagons, obtained by rotating and shrinking the faces of a regular dodecahedron, together with 20 hexagons centered at the vertices of the dodecahedron.*

A slightly more complicated example is S_4. It is generated by three elements $a = (12)$, $b = (23)$, and $c = (34)$. A presentation is

$$S_4 = \langle a, b, c | a^2 = b^2 = c^2 = 1, (ab)^3 = (bc)^3 = (ac)^2 = 1 \rangle .$$

To construct its graph, first make some copies of the ab hexagon for the S_3 subgroup generated by a and b, and similarly make some copies of bc hexagons. The subgroup generated by a and c is $\mathbf{Z}_2 \times \mathbf{Z}_2$, and its graph is a square with edges labeled alternately a and c. Make copies also of ac-squares. Take one copy of each polygon, and fit them together around a vertex, gluing an a edge to an a edge, *etc.* Around the perimeter of this figure, keep gluing on a copy of the polygon that fits. If you do this systematically, layer by layer, you will have constructed a polyhedron — it is a truncated octahedron. All the edges from the underlying octahedron are labeled b, while the squares produced by truncating the vertices are labeled $acac$.

The reader may enjoy working out the graph of the alternating group A_5, using generators $a = (12)(34)$, and $b = (12345)$. Note that they satisfy the relations $b^5 = 1$ and $(ab)^3 = (135)^3 = 1$. Try kicking around the construction, with white $ababab$ hexagons and black $bbbbb$ pentagons.

Of course, graphs of groups don't always work out so nicely or so easily, but often, for simple presentations, they can be worked out, and they tend to have a nice geometric flavor.

§4. LOZENGES

We will begin with a relatively easy tiling problem. Suppose we have a plane ruled into equilateral triangles, and a certain region R bounded by a polygon π whose edges are edges of the equilateral triangle network. When can R be tiled by figures, let us call them lozenges, formed from two adjacent equilateral triangles?

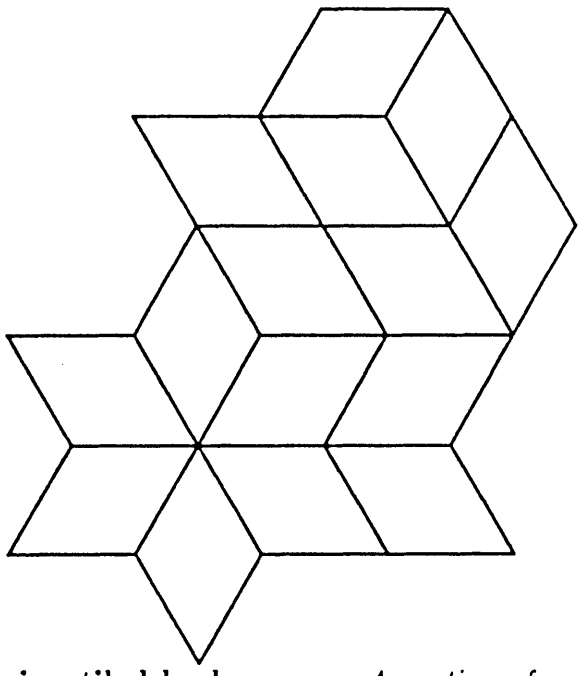

Figure 4.1. A region tiled by lozenges. *A portion of an equilateral triangular subdivision of the plane, tiled by lozenges.*

To analyze this problem, we first establish a labeling convention. We arrange the triangulation of the plane so that one set of edges is parallel to the x-axis, or at 0°. Label these directed edges a, label b the directed edges pointing at 120°, and c the edges pointing at 240°. This labeling is homogeneous, so it is the graph of a group A. We can read off relators for A by tracing out the boundary curves of triangles: A satisfies $abc = 1$ and $cba = 1$. If desired, the first relation could be used to eliminate c; the second relation then says that $ba = ab$. The group A is $\mathbf{Z} + \mathbf{Z}$, as we could have seen anyway by its action on the plane.

The shape of the polygon π is determined by the sequence of edges it traces out; this is a word in the generators a, b, c of A. Rather than thinking of it as a word, we prefer to think of it as an element $\alpha(\pi)$ in the free group F with generators a, b, c. The fact that π closes up is equivalent to the condition that the homomorphism $F \to A$ send $\alpha(\pi)$ to the identity.

If a lozenge is placed in the triangular network, its boundary can be traced by one of three elements, depending on its orientation: that element is either $L_1 = aba^{-1}b^{-1}$, $L_2 = bcb^{-1}c^{-1}$, or $L_3 = cac^{-1}a^{-1}$. The precise word depends on the starting point on the boundary of the lozenge, but starting from a different vertex only changes the word by a circular permuation; the two choices give conjugate elements of F. The *lozenge group* L is defined by these relators, that is

$$L = \langle a, b, c | L_1 = L_2 = L_3 = 1 \rangle .$$

Actually, the three relations say that the three generators commute with each other, so that $L = \mathbf{Z}^3$.

We claim that if the region R can be tiled by lozenges, then the image $I(\pi)$ of $\alpha(\pi)$ in L must be trivial. In fact, suppose that we have such a tiling. If R consists of a single tile, the claim is immediate. Otherwise, find a simple arc in R which cuts R into two tiled subregions R_1 and R_2. By induction, we may assume that $I(\pi_1)$ and $I(\pi_2)$ are both trivial, where π_i is a polygonal curve tracing around ∂R_i. But $I(\pi) = I(\pi_1) * I(\pi_2)$, so $I(\pi)$ is also trivial.

There is a very direct geometric interpretation: think of the graph $\Gamma(L)$ as the 1-skeleton of a cubical tesselation of space, oriented so that cubes are on their corners: more precisely, so that the two endpoints of any path labeled abc are on the same vertical line. The 2-complex $\Gamma^2(L)$ is the union of the faces of the cubes. A lozenge in the plane is the orthogonal projection of a square face of a cube. Given a path π in the plane, arrange it (for notational purposes only) so the base point $*$ lies below the base point 1 of $\Gamma(L)$. Lift it edge by edge to a path in $\Gamma(L)$. When you make a complete circuit around π, you may or may not come back to the starting point in $\Gamma(L)$. The invariant $I(\pi) \in L$ is the ending vertex. This invariant of necessity lies in the kernel of the map $L \to A$, which is isomorphic to $\mathbf{Z}$: it can be described simply as the net rise in height.

If R can be tiled by lozenges, the tiling itself can be lifted, tile by tile, into $\Gamma^2(L)$, that is, into the 2-skeleton of the cubical tesselation. This gives another proof that the invariant $I(\pi)$ must be 1 if R can be tiled. In fact, if you look at a tiling by lozenges, you can imagine it so that it springs out at you in a three-dimensional picture.

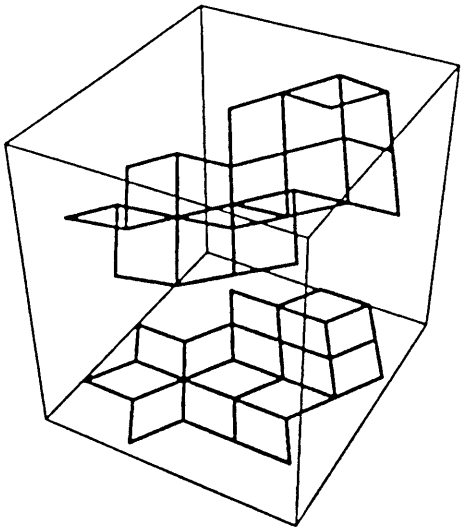

Figure 4.2. Three-dimensional interpretation of lozenge tiling. *If a region R can be tiled by lozenges, then the lozenge pattern lifts to the 2-skeleton of a cubical tiling of R^3, oriented diagonally to the plane of the lozenges.*

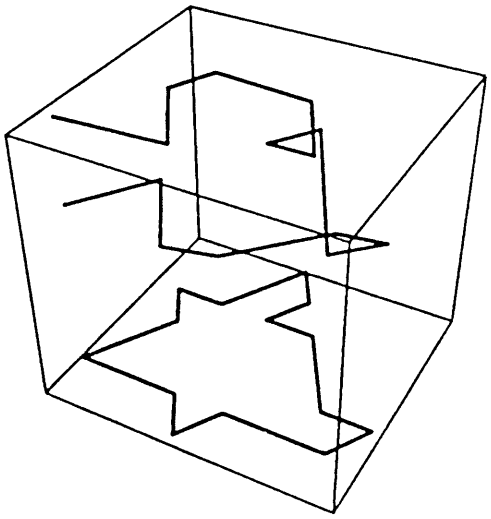

Figure 4.3. Nontileable region. *The region in the plane enclosed by the polygonal curve cannot be tiled by lozenges, since when it is lifted to the cubic network, it fails to close.*

Algebraically, given the word representing π, the net rise in height is simply the sum of the exponents. The condition is that π heads at a bearing of 0°, 120° or 240° the same length of time it heads at a bearing of 60°, 180° or 300°.

This condition can be seen in an alternative way using a coloring argument. The triangles in the plane have an alternating coloring, with *abc* triangles colored white and *cba* triangles colored black. Each lozenge covers one triangle of each color — therefore, if R can be tiled,

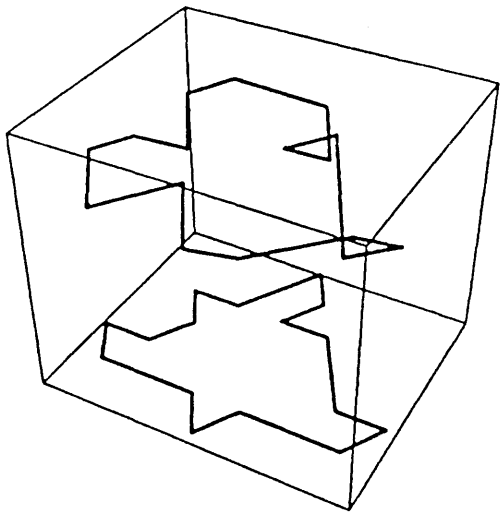

Figure 4.4. Potentially tileable region. *The boundary curve of this region lifts to a closed curve, so it meets the group-theoretic tiling condition. An actual tiling will be shown in 6.1, High lozenge tiling.*

the number of white triangles must equal the number of black triangles. The difference in fact can be shown to be the net rise in height of α, as measured in main diagonals of cubes. The coloring consideration really gives a more elementary derivation that $I(\pi)$ must vanish for a tiling to be possible. However, this and related coloring arguments in general cannot give as much information as $I(\pi)$. One way to think of it is that coloring arguments are the abelian part of the group theory. If the group is abelian as in the present case, or more generally if the subgroup consisting of invariants $I(\pi)$ for *closed* paths is abelian, then that information is sufficient.

The algebraic condition that $I(\pi) = 1$ is not sufficient to guarantee a tiling by lozenges. There are curves π which go around nearly a full circle, with the lift in $\Gamma(\mathcal{L})$ rising considerably, and then instead of closing, they circle around another loop which brings them down to the starting height. If R could be tiled by lozenges, it could be divided into two regions by a fairly short path along edges of lozenges; but the rise in height for one side would be forced to be still positive, which would be a contradiction. We will return later to give a necessary and sufficient condition for a tiling by lozenges, along with a formula for a tiling if such exists.

§5. Tribone tilings

Here is another example, for which other methods seem inadequate. I first heard this problem in an electronic mail inquiry from Carl W. Lee (ms.uky.edu!lee) in Kentucky.

> Last semester, a number of us here became interested in a combinatorial problem that was making the rounds. I'm sure you already have heard of it, and we heard a rumor that John Conway had solved it. It concerned a triangular array of dots. The problem was to pack in as many segments as

possible, where each segment covered three adjacent dots in one of the three directions, and no two segments were allowed to touch. Is there any size configuration that admits a packing such that each dot is covered? Do you know anything about the status of this problem? Thanks in advance.

I hadn't heard of it, but I asked Conway about it. We sat down together, and he worked it out.

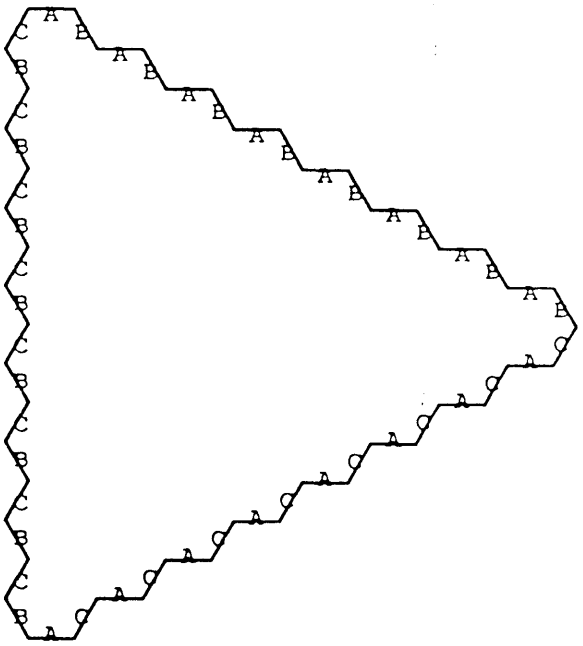

Figure 5.1. Triangle of hexagons. *A triangular array of hexagons, eight on a side. Can this be tiled by tribones?*

This question can be alternately formulated in terms of a triangular array of hexagons. The problem is to show that one cannot tesselate the region using tiles made of three hexagons hooked linearly together. More generally, one can ask for the minimum number of holes left in an attempt to tile the region by these tiles.

If the region has side length n, then the number of hexagons is $n(n+1)/2$. A first, necessary condition is that n or $n+1$ is divisible by 3, that is, n is congruent to 0 or 2 mod 3. Note that if it is ever possible to solve the problem when n is congruent to 2 mod 3, one can extend the solution by adding a row of tiles along one side, to derive a solution for $n+1$.

Label each side in the hexagonal grid with an a, b, or c, according to the direction of the edge: a if it is parallel to the x axis, b if the angle from the x-axis to the edge (measured counterclockwise) is 60°, and c if this angle is 120°. Thus, the sides of every hexagon are labeled *abcabc*.

This labeling gives the 1-skeleton of the grid the structure of a group graph, where the group is

$$A = \langle a, b, c | a^2 = b^2 = c^2 = (abc)^2 = 1 \rangle .$$

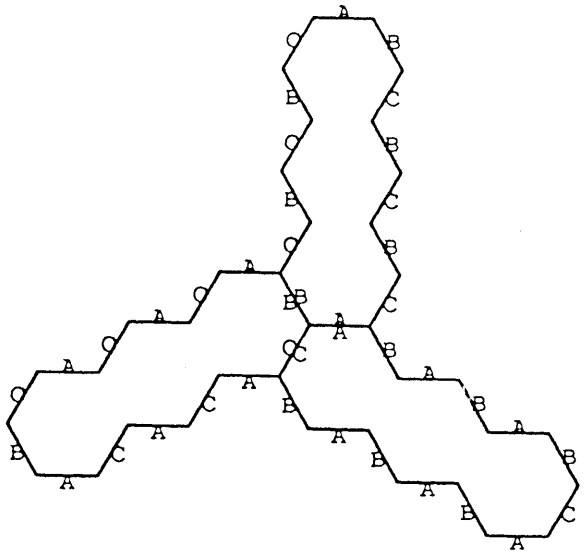

Figure 5.2. Tribones in three orientations. *There are three possible orientations for a tribone, in an array of hexagons. With our labeling convention, they are labeled in three different ways.*

The group is a group of isometries of the plane, generated by 180° revolutions about the centers of the edges; it also contains the 180° revolutions about the centers of the hexagons. The group A is sometimes called the $(2,2,2,2)$-group.

A path π in the 1-skeleton of the hexagonal grid now is determined by a word in the generators of A. We prefer to think of this in a slightly different way: π determines an element $\alpha(\pi)$ in the free product $F = \mathbf{Z}_2 * \mathbf{Z}_2 * \mathbf{Z}_2$. We are particularly interested in closed paths, that is, elements of the kernel of $F \to A$. Unfortunately, this kernel is infinitely generated: it is a free group whose generators are given by arbitrary paths p_1, followed by a circuit around one of the three hexagons at the endpoint of p_1, followed by the p_1^{-1}.

The standard tile, let us call it a *tribone*, can be laid in the plane in three different orientations. Circuits around the tribones in these three orientations trace out the elements

$$T_1 = (ab)^3 c (ab)^3 c$$
$$T_2 = (bc)^3 a (bc)^3 a$$
$$T_3 = (ca)^3 b (ca)^3 b.$$

If π is a simple closed circuit in the plane such that the region R bounded by π can be tiled by these tribones, then the image $I(\pi)$ of $\alpha(\pi)$ in the tribone group

$$T = \langle a, b, c | a^2 = b^2 = c^2 = T_1 = T_2 = T_3 = 1 \rangle$$

must be trivial.

The relation T_1 says that c conjugates $(ab)^3$ to its inverse. Observe that a and b also conjugate $(ab)^3$ to its inverse – in fact, this is already true in F. In other words, $(ab)^3$ generates a normal subgroup, and it commutes with every word of even length. Similarly,

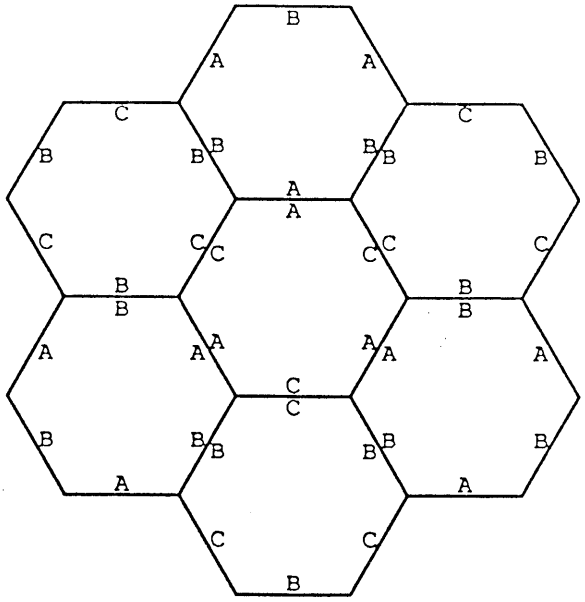

Figure 5.3. Second hexagonal group. *The group T_0 also has a graph isomorphic to the edges of a hexagonal tiling of the plane.*

$(bc)^3$ and $(ca)^3$ generate normal subgroups. Together, the three elements generate a normal abelian subgroup J of T.

To form a picture of T, let us first look at the quotient group $T_0 = T/J = \langle a, b, c | a^2 = b^2 = c^2 = (ab)^3 = (bc)^3 = (ca)^3 = 1 \rangle$. The graph of T_0 can readily be constructed: take an infinite collection of three types of hexagons, with their edges labeled by the relations C_1, C_2 and C_3. These glue together to form a hexagonal pattern in the plane, where each vertex has one a edge, one b edge, and one c edge incident to it. The group T_0 acts faithfully as a group of isometries of the plane, generated by reflections in the edges of this hexagonal tiling: it is a triangle group. It is curious that even though the groups A and T_0 and the labeled graphs $\Gamma(A)$ and $\Gamma(T_0)$ are different, when the labels are stripped they become isomorphic.

If the region R can be tiled by tribones, then $\alpha(\pi)$ must map to the trivial element of T, so it maps to the trivial element of T_0. In our case, the region is a triangular array of hexagons, and its boundary can be taken as $\alpha(\pi) = (ab)^n(ca)^n(bc)^n$.

Obviously, if n is a multiple of 3, the image $I(\pi)/J$ in T_0 is trivial. In the other case, that n is 2 more than a multiple of 3, it is also trivial. This is easily seen by tracing out the curve in our array of hexagons, or by noticing that one can add additional tribones along one edge to form a triangular region with side length $n+1$, which is a multiple of 3. Since we have pushed π only across tribones, $I(\pi)$ is the same for the two cases.

Since T_0 was not sufficient to detect the nontriviality of $I(\pi)$, we need to finish our job, and build a picture of T. First, look at the path in the graph of T_0 determined by the element T_1. Start at a vertex $*$ where the circuit $C_1 = ababab$ goes counterclockwise around a hexagon. Then T_1 goes counterclockwise around this hexagon, then along the c edge, clockwise around the C_1 hexagon through that vertex, and back along the c edge to

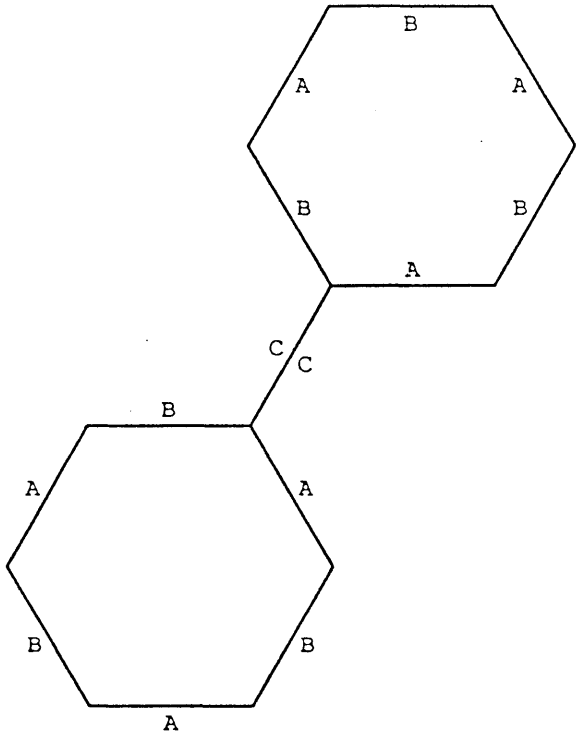

Figure 5.4. Alternate image of tribone. *By construction, the tribone relations are satisfied in the groups T and hence $T_0 = T/J$. This is the image of one of the tribone relators in the graph of the group T_0. Note how it encircles two AB-hexagons, once clockwise and once counterclockwise.*

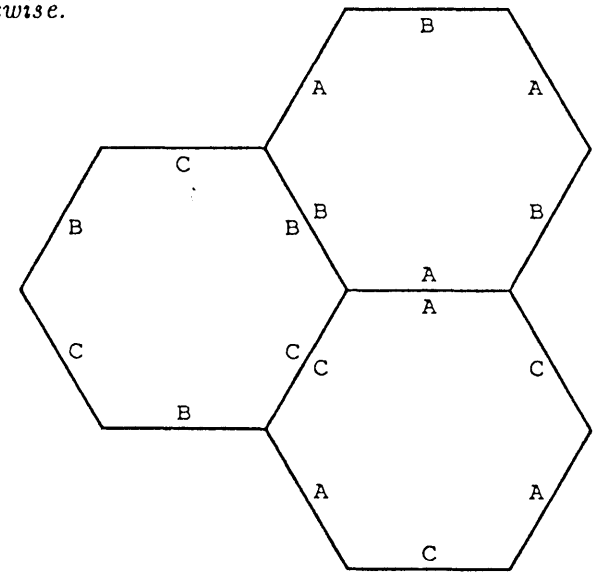

Figure 5.5. Alternate image of a triangle. *The triangle word $(ab)^n(ca)^n(bc)^n$ of size $n = 3m$ or $n = 3m + 2$ maps to the trivial element in T_0. In the diagram above, if $n = 3m$, trace the word starting at the center. If $n = 3m + 2$, start b from the center.*

close. In particular, the signed total of C_1-hexagons enclosed (counted according to degree

of winding with counterclockwise circuits counted positively), is 0.

It is not hard to describe now the full group T, which is an extension of the form $J = \mathbf{Z}^3 \to T \to T_0$. We can interpret an element of T to be a vertex v in the graph of T_0, together with a path p from $*$ to v, subject to the equivalence relation that if q is another path from $*$ to v, then $p \sim q$ if the signed totals of C_1, C_2, and C_3 hexagons are all 0. (Of course, if we pick one path such as p from $*$ to v, then other paths from $*$ to v are determined by three arbitrary integers, which specify these signed totals.) With this definition, the relations T_i are obviously satisfied, hence the group so constructed is at least a quotient group of T. But we have already seen that the kernel J of the map $T \to T_0$ is abelian, and generated by C_i. In the construction, this kernel is the free abelian group on the C_i, so it must in fact give T.

Once we know T, we can read $I(\pi)$ by inspection. As we saw, it suffices to consider the case $n = 3k$; the invariant is $C_1^k C_2^k C_3^k$, which is obviously not 1, so the tiling is impossible.

One can ask whether this method gives a lower bound on the number of holes one is forced to leave, in a partial tiling of R by tribones. To study this question, we should examine the subgroup K of T generated by elements of the form $I(\gamma)$, where γ is a path in the graph of A going from $*$ to some point v, circumnavigating a hexagon, and returning. In other words, K is the kernel of the map $T \to A$. Note that $\alpha(\gamma)$ has the form $gabcabcg^{-1}$, where g is arbitrary. In the group T_0, $abcabc$ acts as a translation. The conjugates of $abcabc$ in T_0 are translations in three different directions spaced at 120° angles, and the subgroup they generate is isomorphic to $\mathbf{Z}^2$. In K, there are actually an infinite number of different conjugates of $abcabc$: if g acts as a translation in T_0, then the commutator $gabcabcg^{-1}cbacba$ is trivial in T_0, but it might not be trivial in T: this path may enclose an arbitrary number m of hexagons of type C_1, and an equal number of type C_2 and C_3.

The subgroup K is therefore a nilpotent group, generated by $s = abcabc$, $t = bcabca$, and $u = C_1C_2C_3$, with presentation

$$K = \langle s, t, u | [s, u] = [t, u] = 1, [s, t] = u^3 \rangle .$$

It is easy to check that every element of K is realized as $I(\pi)$, for some simple closed curve π in the plane.

Even though the invariants associated with triangular regions take larger and larger values in I, this does not give any information limiting the number of holes: for instance, three holes $g_i abcabcg_i^{-1}$ can yield u^k, for arbitrarily high k. In fact, it is possible to tesselate the triangular region of size n with tribones except for 1 hole, if $n \equiv 1(3)$, by placing the hole exactly in the middle, and then arranging concentric triangular layers of tribones around this hole. From these examples, tribone tilings with 3 holes are easily constructed when $n \equiv 0(3)$ or $2(3)$. It does give some information, however: in the case that $n \equiv 2(3)$ or $n \equiv 0(3)$, the conjugacy class changes ("increases") with n, which implies that the length of the minimum closed loop enclosing all the holes has to go to infinity with n. In the case $n \equiv 1(3)$, the conjugacy class of $I(\pi)$ is constant — since the region can always be tiled with a single hexagon missing, $I(\pi)$ is conjugate to $abcabc$. However, the actual word changes with n, which implies that the missing hole cannot be too close to the boundary. Perhaps a careful analysis would show that if there is a single hole, it must be exactly in the center of the triangle.

§6. Dominoes and lozenges revisited

Conway's tiling groups are quite versatile, provided you can work out the group determined by the tiles. Even when (or perhaps especially when) the invariant $I(\pi)$ gives no information which could not have been easily obtained by other means, the geometric picture of the graph of the group can sometimes be exploited to give not just an algebraic criterion, but a precise geometric criterion for the existence of a tiling.

When G is a tiling group (with presentation given by a set of tiles), we define a measure of area in $\Gamma^2(G)$ to be the area defined by projection to the plane: the area of a 2-cell is the area of a corresponding tile. When the algebraic invariant $I(\pi)$ is 1, the curve π bounding R lifts to a closed $\tilde{\pi}$ in $\Gamma(G)$. We can ask, what is the minimum area of a surface S in $\Gamma^2(G)$ with boundary $\tilde{\pi}$? This area is necessarily at least as great as the area of R. If it is equal, then the images of the 2-cells of S must be disjoint, so that they form a tiling of R. There are several approaches which are sometimes successful for calculating this minimal area, but there is one particular situation when there is a really definitive solution: when $\Gamma^2(G)$ can be enlarged, by adding 3-cells, to make a contractible 3-manifold. In this situation, there is a "max flow min cut" principle which guarantees an efficient algorithm for finding a minimal surface.

Rather than going on with the general theory, we will illustrate this with two examples.

First we revisit the lozenge question.

If R is a union of triangles in the plane, and if v and w are vertices in R, possibly on the boundary, define $d(v,w)$ to be the minimum length of a positively directed edge-path in R (possibly going on the boundary) joining v to w. This "distance" function d is not symmetric, since we cannot simply reverse an edge path. Any closed positively directed edge path has length a multiple of 3, so the $d(v,w)$ is defined modulo 3 independent of path. The three vertices of a triangle take the three distinct values modulo 3. If R is connected, it is always possible to find at least one positively directed path from v to w, so $d(v,w)$ is well-defined.

Consider the lifting of any tiling of R by lozenges to the cubical network, $\Gamma^2(L)$. This is determined by a height function $h(v)$ for the vertices v. We can choose the vertical scale so that h is integer-valued, and each edge of a lifted lozenge increases in height by 1; the edge of the triangular network covered by the lozenge lifts to a diagonal of a square, and decreases in height by 2. It follows that $h(w) - h(v) \geq d(v,w)$.

The boundary path π determines a unique height function h on its vertices, up to constants. This gives a necessary condition that R can be tiled: for any two vertices v and w on π, $h(w) - h(v) \geq d(v,w)$.

If π satisfies this necessary condition, then there is a unique maximally high lozenge tiling: define

$$h(x) = \min_{v \in \pi} \{d(v,x)\}.$$

To produce the actual tiling, place a lozenge so as to cover an edge where the height changes by 2. Since the three vertices of a triangle take distinct values modulo 3, and since h increases by at most 1 along any edge, each triangle has exactly one edge where h changes by 2: therefore, the collection of lozenges is a tiling.

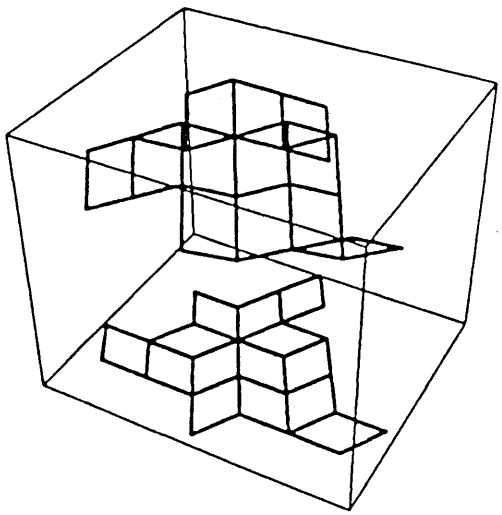

Figure 6.1. High lozenge tiling. *The "highest" lozenge tiling compatible with the boundary curve.*

There is a simple algorithm for quickly computing h, and the tiling: rather than spell it out, we will describe the analogous algorithm for dominoes.

A closed path π in a square grid can be described by an element $\alpha(\pi)$ of the free group $F(x,y)$, which maps to the trivial element of the $A = \mathbf{Z}^2$. If the region R bounded by π can be filled with dominoes, then the image $I(\pi)$ of $\alpha(\pi)$ in the domino group

$$G = \langle x, y | xy^2 = y^2x, yx^2 = x^2y \rangle$$

must be trivial.

What does the graph of G look like? We can construct a picture in $\mathbf{R}^3$, as follows. Fill the xy-plane with a black and white checkerboard pattern. Above the black square $[0,1] \times [0,1]$, construct a right-handed helix, joining $(0,0,0)$ by a line segment to $(0,1,1)$, to $(1,1,2)$, $(0,1,3)$, $(0,0,4)$, and so on: the x and y coordinates here marching forever around the boundary of the square, while the z coordinate increases by 1 each move. Similarly, $(0,0,0)$ is connected to $(0,1,-1)$, *etc.* Construct a similar helix above each black square. Label each edge x or y, according to to its image in the plane. Note that this creates left-handed helices above the white squares. The boundary of any domino in the plane lifts to a closed path in this graph we have constructed. Since the graph has a simply-transitive group of isometries, it is the graph of a group. Since it satisfies the domino relations, it is at least a quotient group of the domino group G. It is not hard (and strictly speaking, it is not logically necessary) to verify that this graph is indeed the graph of G.

The curve π lifts to a curve $\bar{\pi}$ in the graph of G. A convenient way to denote this, in the plane, is to record the height of the lift next to each vertex of π in the plane. The rule is simple: one can start with 0 at some arbitrary vertex. Along any edge of π which has a black square to its left, the height increases by 1. Along any edge with a white square to its

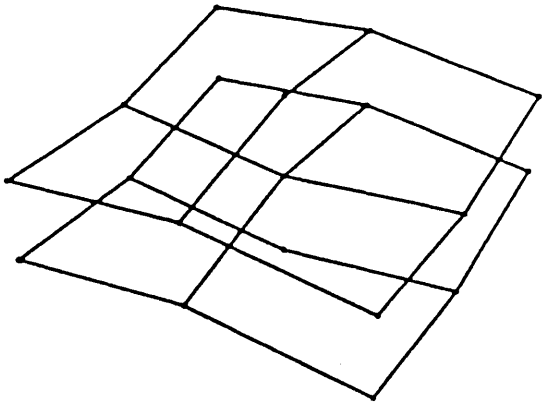

Figure 6.2. The domino group. *The graph of the domino group is a union of square helices over the squares of a checkerboard, alternating in handedness. A domino anywhere in the plane lifts to this graph, starting at any point. This illustration shows two coils of four neighboring helices.*

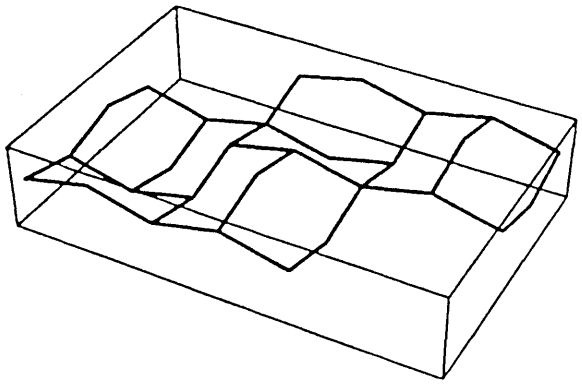

Figure 6.3. Domino tiling. *A tiling by 9 dominoes, lifted to the graph of the domino group.*

left, the height decreases by 1. A necessary condition that R can be filled with dominoes is that the height after traversing once around the curve is 0.

There is a criterion and construction for a domino tiling, analogous to the construction for lozenges. Here is how the formula can be worked out, on a sheet of grid paper. Begin, as above, by labeling the height of each vertex of π. The heights consist of the integers in some interval, $[n, m]$. We will construct a height function on all vertices of R, beginning with $n + 1$, and working up. Suppose, inductively, that we have finished with all vertices of height less than or equal to k. For each vertex v of height k, and for each edge e leading

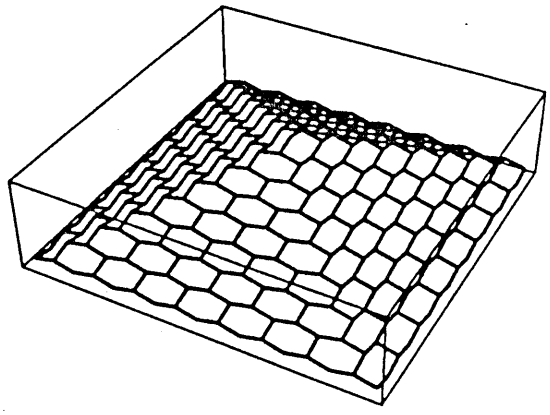

Figure 6.4. Domino roof. *This is the tiling which the algorithm yields, when applied to a* 16×16 *square grid. This is the tiling which has the highest lifting to the graph of the domino group of any tiling by dominoes.*

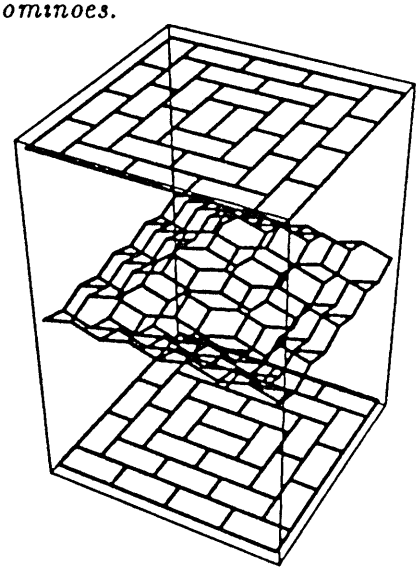

Figure 6.5. Domino bubble. *This illustration shows both the highest and the lowest tiling by dominoes of a standard checkerboard. They are isomorphic, differing only by a* $90°$ *rotation of the checkerboard (interchanging colors). The upper tiling is shown in the upper plane as well as the upper surface of the bubble, the lower tiling in the lower plane and the lower surface of the bubble. The bubble they form encloses the lift of any tiling by dominoes. Possible tilings are 'like' Lipschitz functions in the square with Lipschitz constant 1, as measured in the Manhattan metric. The limits of domino tilings, lifted to the graph of the group, as the grid size goes to zero, are exactly such Lipschitz functions.*

from v which has a black square on its left, consider the second endpoint w of e. If the

height of w has been previously defined, and if it is not greater than $k+1$ leave it as is. If the height is defined and greater than $k+1$, then a domino tiling is impossible: give up. Otherwise, define the height of w to be $k+1$.

If this procedure reaches a successful conclusion, each edge of R has a difference of heights of its two endpoints of either 1 or 3. (Note that the height modulo 4 is determined by the point in the plane.) Erase all the edges whose endpoints have a difference of height of 3. What is left is a picture of a tiling by dominoes.

§7. Triangles

Here is a related sequence of tiling problems which are resistant to direct attempts at general solution, but translate nicely into the realm of group theory.

Consider, again, a triangular array of dots, with N dots on each side. Is it possible to subdivide this array into disjoint triangular arrays of dots with M on each side? We suggest the reader indulge in experimentation with a few cases, before reading further. For example, the cases $M = 2$ with N ranging from 2 to 12 are interesting.

...

As in the case of the tribones, this translates into a tiling problem: given a triangular array of hexagons with N hexagons per side, can one tile it by tiles T_M which are triangular arrays of hexagons M per side? We can express this with notation as in the case of tribones: label the edges of the underlying hexagonal tiling by a's, b's, and c's. Given a path π in the plane, it is described by an element $\alpha(\pi)$ of $F = \langle a, b, c | a^2 = b^2 = c^2 = 1 \rangle$. If the region R bounded by π can be tiled by the copies of T_M, then the image $I(\pi)$ of $\alpha(\pi)$ is trivial in the group

$$G_M = \langle a, b, c | a^2 = b^2 = c^2 = 1, t_M = 1 \rangle,$$

where t_M represents the boundary curve of the tile T_M,

$$t_M = (ab)^M (ca)^M (bc)^M.$$

A parallelogram of hexagons with M hexagons on one side and $M+1$ on the other can be tiled by two copies of T_M. This implies that $(ab)^M$ commutes with $(bc)^{M+1}$ and with $(ca)^{M+1}$, and so forth.

These relations imply that $(ab)^M$ commutes with $(bc)^{M(M+1)}$, and they also imply that $(ab)^{M+1}$ commutes with $(bc)^{M(M+1)}$. Combining these two facts, it follows that (ab) commutes with $(bc)^{M(M+1)}$. Geometrically, one can tile an $M \times M(M+1)$ parallelogram and an $(M+1) \times M(M+1)$ parallelogram. Their difference is a $1 \times M(M+1)$ parallelogram: this can be tiled in a certain algebraic sense as the difference of the two.

It will simplify the picture at this point if we pass to the subgroups F^e and G_M^e generated by words of even length. Since all relations have even length, the wordlength modulo 2 describes a homomorphism of F and G_M to $\mathbf{Z}_2$, and these subgroups have index 2. The group F^e is the free group on 2 generators, but a more symmetric description is

$$F^e = \langle x, y, z | xyz = 1 \rangle,$$

where $x = ab$, $y = bc$, and $z = ca$. A presentation for the group G^e_M is obtained by adjoining relations coming from t_M to F^e: it requires two relations, one obtained by transcribing t_M directly, and the other transcribing the conjugate of t_M by an element of odd length. Using $t_M = 1$ and $bt_Mb = 1$, we obtain

$$G^e_M = \left\langle x, y, z | xyz = 1, x^M y^M z^M = 1, x^{-(M+1)} y^{-(M+1)} z^{-(M+1)} = 1 \right\rangle.$$

G^e_M has an interesting alternate generating set: $X = x^M$, $X' = x^{-(M+1)}$, together with Y, Y', Z and Z' defined similarly, clearly generate. We have already seen that X, Y, and Z commute with X', Y' and Z'.

The elements $s = X^{M+1}$, $t = Y^{M+1}$, and $u = Z^{M+1}$ commute with everything in G^e_M, so they generate a central subgroup J which is $\mathbf{Z}^3$ or a quotient. Let $G^0_M = G^e_M/J$. We will analyze the structure of G^0_M, and from that construct G^e_M.

In G^0_M, X, Y, and Z satisfy relations

$$XYZ = 1, X^{M+1} = Y^{M+1} = Z^{M+1} = 1.$$

These relations describe the orientation-preserving $(M+1, M+1, M+1)$ triangle group, which acts as a discrete group of isometries on the Euclidean plane if $M = 2$ and on the hyperbolic plane if $M > 2$. We have not checked that these generate *all* the relations on X, Y, and Z, but we immediately deduce that the subgroup H of G^0_M generated by X, Y and Z is a quotient of this triangle group. But there is a homomorphism f of the original group G_M to the full triangle group (including reflections), defined by sending a, b, and c to reflections in the sides of a $\pi/(M+1), \pi/(M+1), \pi/(M+1)$ triangle. The relation $t_M = 1$ is satisfied, since in this group $(ab)^M = ba$ so that $(ab)^M(ca)^M(bc)^M = (ba)(ac)(cb) = 1$. Note that f sends X to ba, Y to ac and Z to cb, that is, to the standard generators of the $(M+1, M+1, M+1)$ triangle group, and it sends s, t and u to 0. Therefore, H is isomorphic to the orientable $(M+1, M+1, M+1)$ triangle group.

A similar analysis shows that the subgroup H' generated by X', Y' and Z' is the orientable (M, M, M) triangle group. This group acts on the sphere, the Euclidean plane, or the hyperbolic plane when $M = 2$, $M = 3$, or $M \geq 4$. The analogous homomorphism f' maps G_M to the full (M, M, M) triangle group, mapping a, b, and c to the standard generators.

The two subgroups H and H' intersect trivially (as seen from the effects of f and f'), they generate G^0_M, and they commute with each other. Therefore, G^0_M is the product $H \times H'$ of the two triangle groups.

Now we need to determine the kernel J of the quotient $G^e_M \to G^0_M$, and the structure of the central extension. As in the tribone case, we can do this geometrically, in terms of areas enclosed by curves. The graph Γ of the full $(M+1, M+1, M+1)$ triangle group is formed from copies of three kinds of $2(M+1)$-gons, with perimeters labeled $(ab)^M$, $(ca)^M$ and $(bc)^M$, with one of each kind meeting at each vertex. Arrange the orientation so that 1 is an "even' vertex that is, the a, b, and c edges emanating from 1 are in counterclockwise order. Then the relation t_M based at v encloses positively one copy of each type of polygon, while the conjugate bt_Mb encloses negatively one copy of each type of polygon.

Similarly, the graph Γ' of the full (M, M, M) triangle group is made from three kinds of $2M$-gons. Starting at the 1, which we suppose is an even vertex, the relation t_M encloses positively one copy of each type of polygon, while bt_Mb encloses negatively one copy of each. However, in the case $M = 2$, there the entire graph is finite: it is the 1-skeleton of a cube, and the number of polygons enclosed by a curve is well-defined only modulo 2.

First let's deal with the case $M > 2$. We can define an extension K of G_M^0 as an equivalence relation on elements of F^e, as follows. An element g of F^e determines paths $p(g)$ in Γ and $p'(g)$ in Γ'. We define g to be equivalent to h if $p(g)$ ends at the same point as $p(h)$, $p'(g)$ ends at the same point as $p'(h)$, and if the closed loop $p(g)p^{-1}(h)$ encloses the same numbers of ab-polygons, bc-polygons, and ca-polygons as $p'(g)p'^{-1}(h)$.

In particular, an element of the kernel of the map of K to $H \times H'$ maps to closed loops in both pictures, and is determined by the triple of differences of the number of polygons enclosed. The elements s, t and u map to $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. It follows that $K = G_M^e$, and $J = \mathbf{Z}^3$ (provided $M > 2$.)

The boundary of the size N triangle T_N can be described by the element $t_N = (ab)^N(ca)^N(bc)^N$. The path $p(t_N)$ in Γ closes only when N is 0 or -1 mod $M+1$, while the path $p'(t_N)$ closes only when N is 0 or -1 mod M. Since M and $M+1$ are relatively prime, there are four solutions modulo $M(M+1)$: $0, M, M^2-1, -1$. For values of N satisfying one of these congruence condition, the invariant in G_M^0 is 0, so the invariant is in J; it is a positive multiple of $(1,1,1)$ in all but the trivial case $N = M$.

THEOREM (CONWAY). *When $N > M > 2$, the triangular array T_N of hexagons cannot be tiled by T_M's.*

This analysis has an interesting variation case $M = 2$. Given two elements g and h of F^e, we can define them to be equivalent if $p(g)$ and $p(h)$ have the same endpoints, $p'(g)$ and $p'(h)$ have the same endpoints, and if the numbers of polygons of the three types enclosed by the path $p(g)p(h)^{-1}$ is a multiple k of $(1,1,1)$ which has the same parity as the number of polygons enclosed by $p'(g)p'(h)^{-}1$. This defines a central extension of $H \times H'$ by $\mathbf{Z}^3$ modulo the subgroup generated by $s^2t^2u^2 = 1$. To justify that this group is in fact G_2^e, we must prove that $s^2t^2u^2 = (ab)^{12}(ca)^{12}(bc)^{12} = 1$ in this group, or even better, that it is possible to tile T_{12}. Such a tiling can be found fairly easily — see figure 5.1, the 12-stack by 2-stacks.

The computation of the mod 2 invariant for tilings by T_2's can be rather annoying when done directly. However, there is a neat trick, which enables one to see this invariant geometrically: most regions which have a multiple of 3 hexagons can be tiled easily by T_2's along with tribones. The boundary *abababcababababc* of a tribone maps to closed paths in both Γ and Γ'. In Γ, it encloses a net of 0 of each type of hexagon, as we saw before. In Γ', this curve winds counterclockwise 1.5 revolutions about an ab-face of the cube, goes down a c-edge to the opposite face, winds 1.5 revolutions counterclockwise (with respect to the orientation of the square), and goes up again to close. It is therefore equivalent, in terms of which kinds of squares it encloses, to *abcabc*, which is an odd multiple of $(1,1,1)$.

Therefore, if a region can be tiled with a collection of T_2's together with an odd number of tribones, it cannot be tiled with T_2's. For $0 < N < 12$, only for the values $2, 3, 5, 6, 8, 9, 11$

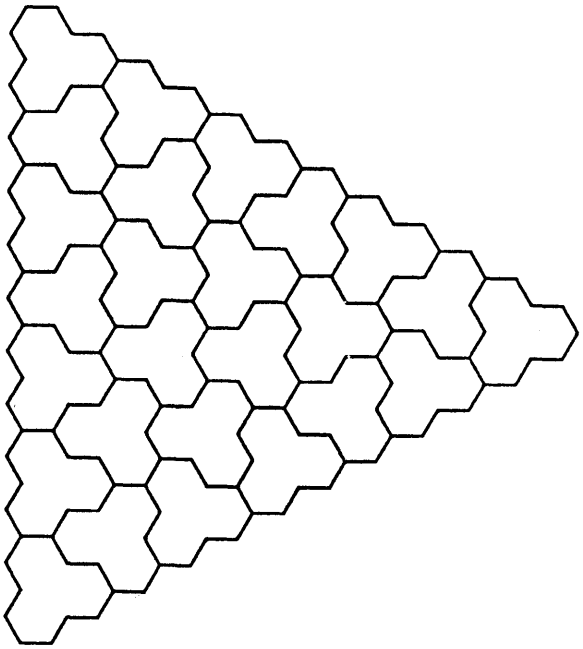

Figure 7.1. The 12-stack by 2-stacks. *The triangle T_{12} can be tiled by T_2's.*

is the number of tiles a multiple of 3. One quickly finds that in the cases T_3, T_5, T_6 and T_8 there is a tiling by one tribone and the rest T_2's, while T_2, T_9, and T_{11} can be tiled.

Given any tiling or partial tiling of T_k, with $k > 1$, it can be extended to a tiling or partial tiling of T_{k+12} by adding a $12 \times k$ parallelogram, together with a T_{12}. The $12 \times k$ parallelogram can be tiled by subdividing into 2×6 and 3×6 parallelograms.

THEOREM (CONWAY). *A triangular array T_k of hexagons can be tiled by T_2's if and only if k is congruent to* 0, 2, 9, *or* 11 *modulo* 12.

§8. SELF-SIMILAR TILINGS

As previously remarked, there are many ways to tile the plane, or to tile $\mathbf{R}^n$. In fact, there is such a variety of tilings of the plane, even by translates of a finite number of polygons, that the question of whether a given set of tiles will tile the plane is undecidable. For any Turing machine, it is possible to construct a finite set of tiles such that these tiles fail to tile if and only if the Turing machine eventually comes to a halt. The output of the Turing machine is recorded, in terms of the tiling, by the tiles at a certain sequence of spots (spaced according to a geometric progression).

Thus it is interesting to force more conditions on a tiling, and see what happens. One interesting theory comes about by examining self-similar tilings: tilings for which each model tile has a subdivision into subtiles, such that when this subdivision is performed on all tiles simultaneously, the resulting tiling is isomorphic, by a similarity of the plane (or space), to the original.

Eventually we will prove a rather general thereom, in which we characterize the set of similarities for selfsimilar tilings of the plane (or of higher-dimensional spaces.) This is

closely akin to the constructions for Markov partitions in dynamical systems. However, what is also interesting about this subject is the particular constructions — at issue is how simple and how nice can self-similar tilings be. The general constructions be very loose, and yield estimates for immense numbers of tiles (or elements of a Markov partition.) Therefore, we will take the time time to discuss several particular examples and constructions.

Pieces of this material are to be found in a number of sources, and I have not been organized enough to work out the appropriate attribution — this writing should be thought of as semi-expository, and many things I say here are not original with me. The theory of Markov partitions underlies most of this, and there is a very extensive mathematical literature. Some of the selfsimilar tilings can be derived from Anosov maps of the torus to itself; V. Arnold developed some connections along this line, particularly for the Penrose tilings. The theory of exotic number bases has a literature with which I am not very familiar; in particular, though, it is discussed in Knuth's Art of computer programming.

Rick Kenyon, currently a graduate student at Princeton, has worked out some beautiful additional constructions for self-similar tilings which I will not discuss here.

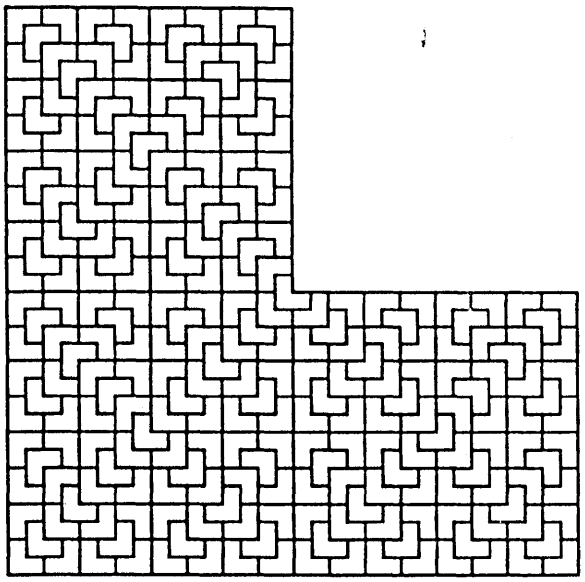

Figure 8.1. Tiling of the plane by trionimoes. *This is a portion of a non-periodic self-similar tiling of the plane by L-shaped trionimoes. The expansion factor is* 2. *The trionimoes come in* 4 *orientations.*

There are rather trivial examples of this phenomenon: for instance, the tiling of the plane by squares is self-similar, with the subdivision rule that each tile subdivides into 4 subsquares.

A slightly more complicated example is an L-shaped trionimo, made from three squares glued together. It can be subdivided into 4 similar figures of half the size. If you expand this shape, and then subdivide again, and continue indefinitely, the limit of this process yields a self-similar tiling of the plane by L-trionimo's

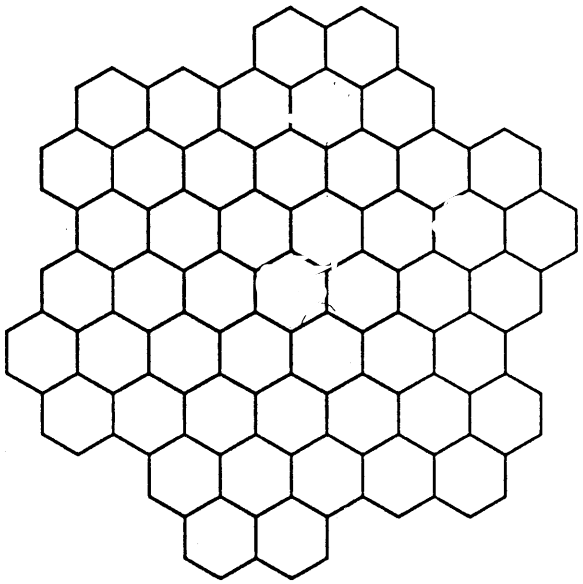

Figure 8.2. Second stage in hexagonal fractal tiles. *The array of hexagons formed by repeatedly seven-coloring, then regrouping the tiles around the blue tiles: second level*

This same process does not quite work with hexagons. A hexagon, together with its 6 neighbors, looks roughly like a hexagon — but not exactly.

Let's modify the shape a bit until it works. As a second approximation to a self-similar tiling, instead of a hexagon, let's use a hexagon together with its six immediate neighbors as a tiling of $\mathbf{R}^2$. These tiles can be used to tile the plane in a hexagonal pattern: to do it, use a periodic seven-coloring of the hexagonal tiling. If one of the colors is blue, then each tile is either blue, or touches exactly one blue tile. Use the clusters centered at blue tiles to tile the plane.

To continue the process, it helps to renormalize the new tiling, so that the lattice of center points of the blue tiles is mapped to the lattice of center points of all tiles. Thus we get a seven coloring of the new tiling of the plane. Group the new tiles by sevens, and renormalize. This process, iterated, converges to a tile of a certain fractal shape. The limiting shape is homeomorphic to a disk, and it tiles the plane in the same combinatorial pattern as the original hexagonal tiling — but now the tiling is self-similar. When we transform the plane, considered as $\mathbf{C}$, by the transformation $z \to \alpha z$ where $\alpha = (5 + \sqrt{-3})/2$, then the image of any tile in the limit pattern is the union of seven tiles.

This example has the feature that all tiles are congruent, so there is only one rule for subdivision. There is a rich collection of examples which can be constructed similarly, but there are even more tilings which have several tile types. We will construct a first example on the real line. The simplest possibility is that there are only two types of tiles, which are intervals: let us call them A and B. We specify that when an A tile is enlarged, it subdivides into an A and a B, and when a B tile is enlarged, it becomes an A. Then the second subdivision of A consists of ABA, which goes to $ABAAB \to ABAABABA \to ABAABABAABAAB$ *etc.*

The numbers of A-tiles and B-tiles, as A is subdivided, are $(1,0) \to (1,1) \to (2,1) \to$

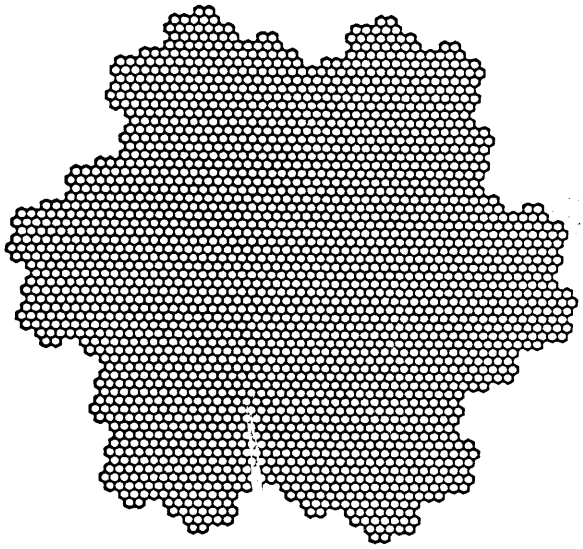

Figure 8.3. Fourth stage in hexagonal fractal tiles. *The array of hexagons formed by repeatedly seven-coloring, then regrouping the tiles around the blue tiles: fourth level*

$(3,2) \to (5,3) \to (8,5) \to \ldots$. They are Fibonnaci numbers. It follows that the expansion constant must be the golden ratio, $\phi = (1+\sqrt{5})/2$. The lengths (a,b) of the two tile types satisfy $\phi a = a + b$ and $\phi b = a$. In other words, (a,b) must be an eigenvector of $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ of eigenvalue ϕ, say $a = \phi$ and $b = 1$. With these choices, we see that the rule for subdivision works.

To construct an actual self-similar tiling with this pattern, we can start an A tile $[0,\phi]$. The expansion of this tile implies that there is a B tile adjacent to it, $[\phi, \phi+1]$. The expansion of the B tile defines another A tile ... and so on. Eventually we get a tiling of the positive real line. The pattern does not actually extend to a strictly self-similar tiling of the negative real line. If we put a B tile $[-1,0]$, it expands and subdivides into a single A, which subdivides back into an AB. We get a pattern which repeats with period 2. If we used instead the rule for the second subdivision, $A \to ABA$ and $B \to AB$, the tiling would be strictly self-similar.

There is a famous 2-dimensional generalization of the preceding example, due to Roger Penrose, and known as the Penrose tiling. The Penrose tiles come in several variants We will describe a version using two shapes of isosceles triangles, the two triangles having side lengths in the golden ratio. To properly define a rule for subdivision, however, we consider the two triangles to come in two types, a left-handed form and a right-handed form. This is graphically represented by putting dark edges on two of the sides of each triangle. In the tiling, dark edges will match with dark edges.

Across any undarkened edge is a mirror image of the given triangle. The union of two such mirror image thin triangles is a *kite*; the union of two mirror image fat triangles is a *dart*.

The rules for subdivision can be applied recursively to get finer and finer subdivisions

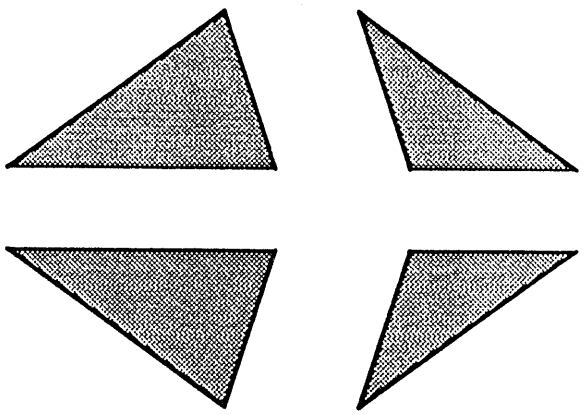

Figure 8.4. Penrose tiles 1. *The four basic triangles, $thin_1$, fat_1, $thin_2$, fat_2, in left-handed and right-handed versions. The handedness is distinguished by the dark edges on two of the sides of each triangle. Dark edges will always match with dark edges, and the tile adjoining along a side without a dark edge will always be the mirror image.*

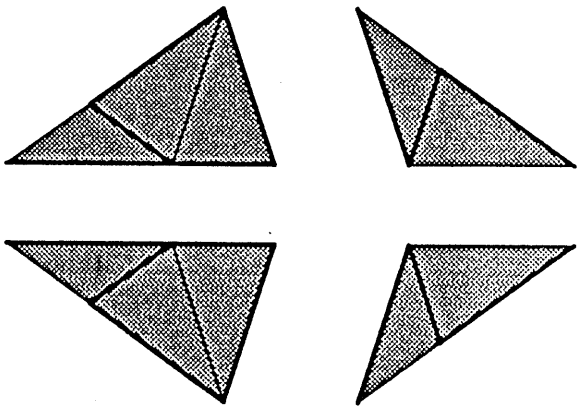

Figure 8.5. Penrose tiles 2. *The rule for subdivision of the four basic triangles. A thin triangle subdivides into two thin triangles and a fat triangle; a fat triangle subdivides into one fat triangle and one thin triangle.*

of a given triangle, or they can be rescaled to give tilings of larger and larger regions in the plane.

To get a self-similar tiling of the entire plane, we can exploit the fact that the subdivision of a right-handed thin triangle contains a right-handed thin triangle. Map the subdivided triangle by a complex affine transformation (a similarity) which sends the small triangle of the subdivision to the original. If this process, subdivision followed by expansion, is

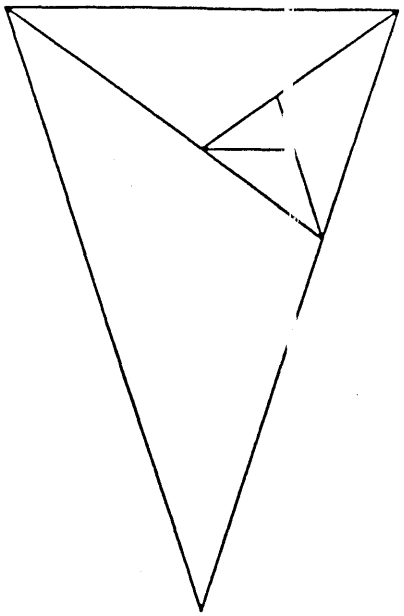

Figure 8.6. rescaling triangles. *Thin triangle rescaled so that the subdivisions of the larger ones match with the smaller.*

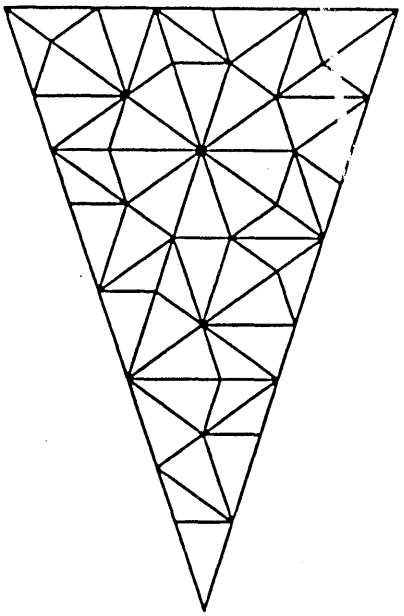

Figure 8.7. expanding subdivisions. *The subdivision of a fourth-generation rescaled triangle.*

iterated, we obtain a tiling of the entire plane.

§9. Solitaire

One construction for self-similar tilings can be described in terms of a kind of game of solitaire. First we will go over the theory for real numbers: it is a nice special case, and

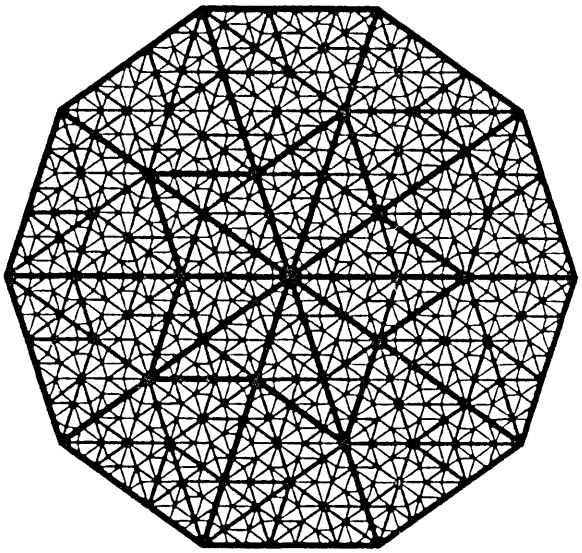

Figure 8.8. Penrose tiles 3. *A self-similar penrose tiling with symmetry the dihedral group of order* 10 *can be constructed by beginning with* 10 *thin triangles of alternating handedness arranged about the origin. The second subdivision of this a configuration contains the original configuration in the center. This picture shows the first* 5 *subdivisions, with the edges of triangles at the various stages having thickness scaled to show the recursive structure.*

enables us to indirectly construct some tilings of the plane.

Consider any real number β greater than 1. There is a canonical way to construct a base β system for the real numbers, which coincides with the usual definition if β is an integer. The definitions are very simple: A not necessarily proper representation in base β means a series

$$d_k\beta^k + d_{k-1}\beta^{k-1} + \cdots + d_0 + d_{-1}\beta^{-1} + \ldots,$$

with $d_i > 0$, also written

$$d_k d_{k-1} \ldots d_0 . d_{-1} \ldots .$$

Such a series need not even converge, but in practice the digits d_i will be bounded, so that it converges to a positive real number which it represents. Improper representations have a lexicographical ordering: they are ordered according to the first digit in which they disagree.

A representation is *strictly proper* if the digits are bounded, and if it is lexicographically the greatest representation of the positive real number it represents. It is *weakly proper* if each finite truncation is *strictly proper*. Thus in base $\beta = 10$, $.999\ldots$ is weakly but not strictly proper, since $1.000\ldots$ is greater in lexicographical order. Improper representations are frowned upon in school, but they have their place. Once, a niece of mine in kindergarten told me she knew what 3 times 11 was: 33.

I asked "Well then, what's 7 times 11?"

"It's seventy-seven."

"Okay, I bet you don't know what 12 times 11 is."

"Twelvety-twelve", she gleefully replied.

We agreed that twelvety-twelve is a perfectly good number, we know what it means, but if we were talking to other people we would tell them one hundred and thirty two — for us, twelvety-twelve is just fine.

The key to understanding any base β is its *carry sequence*, The carry sequence may be described as the sequence of digits of the representation of 1 which is weakly proper, but not strictly proper.

The carry sequence carry(β) may be constructed by a dynamical process, as follows: start with $x = 1$. Repeatedly multiply x by β, and subtract the largest integer d_i strictly less than the result. The sequence d_i so obtained is the carry sequence.

PROPOSITION 9.1. CARRY CHARACTERIZES. *A representation in base β is strictly proper if and only if the sequence of digits starting at any point is lexicographically less than the carry sequence* carry(β).

PROOF: This is pretty obvious. ∎ 9.1, carry characterizes

PROPOSITION 9.2. CARRY SHIFTS LESS. *A sequence of positive integers $\{c_i\}$ is a carry sequence if and only if it has an infinite number of $c_i \neq 0$, and no sequence obtained by dropping a finite number of initial elements is lexicographically greater than it.*

The operation of dropping the first element of a sequence is known as the Bernoulli shift, or one-sided Bernoulli shift.

PROOF: It is clear that the carry sequence for any base β has these properties. Note that if c_i were eventually 0, then x would have arrived at 0 in the dynamical process above, which is impossible.

To prove the converse, consider the map from real numbers to carry sequences. It takes natural order of the real line to lexicographical order.

The set of all sequences of positive integers with a given bound has a natural topology of a Cantor set, with the compact-open topology. If $\{\beta_i\}$ is an increasing convergent sequence of real numbers, then carry(β) also converges to carry(lim($\{\beta_i\}$)). However, carry is discontinous at those β so that the dynamical process above eventually arrives at a discontinuity of the greatest integer function (an integer). On the next step, x is then 1, so that carry(β) is periodic. At such a point, one sees from the dynamical process that if the carry sequence is $.(c_0 c_1 \ldots c_{k-1} c_k)$, the limit from above is $.c_0 \ldots c_{k-1}(c_k + 1)0\,0\,0\ldots$

Consider now the closure of the set of all sequences satisfying the condition of the proposition, that is, all such sequences together with all limits from above at periodic carry sequences. Form the quotient topology, identifying each periodic carry sequence with its limit from above. It is not hard to see that this topological space is homeomorphic to the real line, using the linear ordering. The map $\beta \to$ carry(β) is monotone and continuous in this topology, so by the intermediate value theorem it is surjective.

∎ 9.2, carry shifts less

These observations are closely related to the theory of kneading sequences, [**Milnor Thurston**], which arise in the theory of iterated (non-homeomorphic) maps of intervals.

A *Pisot* number is an algebraic integer such that all its Galois conjugates are strictly inside the unit circle. In more down-to-earth terms, a Pisot number is real number x which is a root of a polynomial $x^n + a_{n-1}x^{n-1} + \cdots + a_0$ with integer coefficients and leading coefficient 1 such that all the roots except x are inside the unit circle in the complex plane.

PROPOSITION 9.3. PISOT CARRY PERIODIC. *The carry sequence for any Pisot base $\beta > 1$ is periodic.*

If $x > 0$ is an element of the field $\mathbf{Q}(\beta)$, then the representation of x in base β is eventually periodic.

Examples. The most famous example is the golden ratio $\phi = 1.618\ldots$. Its carry sequence is .1010101.... This means that a base ϕ representation is weakly proper if and only if the digits are 0's and 1's, and each 1 is followed by at least one 0. It can be seen by computation (or inspection) that the carry sequence for the cubic number $x^3 = x^2 + 1$. $x \approx 1.465571231876768$ is .100100100100...: each 1 must be followed by at least two zeros. Computation shows that the carry sequence for the cubic $x^3 = x + 1$, $x \approx 1.324717957244746$, is .100001000010000 .... From this example one sees that the length of the period can be longer than the degree. The cubic number $x^3 = 3x^2 - 2x + 1$, $x \approx 2.546818276884082079$, is .201111111.... This example shows that the carry sequence can be eventually periodic but not periodic. All these numbers are Pisot numbers.

PROOF: The base-β representation x_β of a positive real number $0 < x < \beta$ is determined by a dynamic process almost identical to the previous: start with x, subtract the greatest integer in x and multiply by β to get the new x.

Even though we have been talking only about the real numbers, somehow multidimensional spaces lurk in the background.

The tensor product $\mathbf{Q}(\beta) \otimes \mathbf{R}$ is a vector V space of dimension d, where d is the degree of β. Multiplication by β extends to an action on V. Each real root of the minimal polynomial for β is an eigenvector for this action, with a 1-dimensional eigenspace, and for each pair of complex conjugate roots there is a 2-dimensional invariant subspace: it has two complex structures, of opposite orientation, in which the action of β is conjugate to multiplication by these complex roots. Since all the characteristic roots but β are inside the unit circle, the dynamics of multiplication by β are to squeeze everything toward the β-eigenspace, and stretch out that eigenspace. Let $S \subset V$ be the hyperplane which is the linear span of the contracting directions, and $U \subset V$ be the expanding subspace.

If x is in $\mathbf{Q}(\beta)$, it defines an element of V, and the dynamic process defining x_β can be interpreted inside V. We repeatedly subtract the largest multiple of $1 \in V$ which keeps x on the same side of S, then multiply by β.

Observe that x always remains in a bounded region of V, as this process is iterated. It can never escape very far from S, since we always guide it back, and it can never escape very far from U, since β squeezes V toward U.

If x starts out as an algebraic integer in $\mathbf{Q}(\beta)$, then it always remains an algebraic integer. The set of all algebraic integers forms a lattice in V, so x can only take a finite

number of values. Therefore, its orbit eventually arrives back at a previous point, and from then on it repeats.

If x is in the field but not an algebraic integer, then there is some integer m such that mx is an algebraic integer. Multiplication by β preserves this property. Therefore the orbit of x remains in the lattice of $(1/m)$ times algebraic integers, and again it must eventually repeat.

9.3, Pisot carry periodic

The converse of this proposition is also true, but it is not true that every algebraic number with an eventually periodic carry sequence is Pisot: numbers with periodic carry sequences are dense, but Pisot numbers form a countable closed subset of $\mathbf{R}$. For a random example, the sequence .(32123012310) satisfies the hypotheses of Proposition 9.2, carry shifts less, so it is the carry sequence of some number β. Simple algebraic manipulation shows that β must be a root of the irreducible polynomial

$$x^{11} - 3x^{10} - 2x^9 - x^8 - 2x^7 - 3x^6 - x^4 - 2x^3 - 3x^2 - x - 1,$$

whose largest root (compare 10.1, Largest integers expand tilings) is

$$3.675589442327944039432480352575683267 4643931\ldots.$$

Computation shows that its carry sequence is indeed

$$.321230123103212301231032123012310321230123\ 032123012310\ldots.$$

This polynomial has two other roots $.340861 \pm .998669i$ (modulus 1.05524) outside the unit circle, so it is not a Pisot number.

One immediately obtains self-similar tilings of $\mathbf{R}$ from any real number with a periodic carry sequence: tile the positive reals according to the 'whole' portion of its base β representation, that is, the portion to the left of the decimal point. Multiplication by β is a shift of the decimal point, so that each tile is taken to a finite union of tiles. The fact that there are only a finite number of different tiles up to congruence is equivalent to the fact that carry(β) is eventually periodic: in fact, the lengths of intervals which occur are exactly the orbit of 1 under the dynamical process for the carry sequence.

It is curious that one also obtains a self-similar tiling of the plane or a higher-dimensional space from this construction, if β is a cubic Pisot number which is also an algebraic unit (this means that the constant term of the minimal polynomial is ± 1). The nice case is when the degree is 3, when we will obtain a self-similar tiling of the plane, a kind of 'Galois conjugate' of the original tiling.

Suppose then that $\beta > 1$ is a Pisot unit. For any algebraic integer $x > 1$ in $\mathbf{Q}(\beta)$, let k be the greatest integer such that $\beta^{-k}x \geq 1$. The sequence of digits for x (shifted k positions) is obtained by the dynamical process starting with $\beta^{-k}x$. Note that since β is an algebraic unit, its inverse is also an algebraic unit, so $\beta^{-k}x$ is an algebraic integer. Therefore, x_β is eventually periodic, terminating in one of a finite set of repeating patterns.

For any algebraic integer $x \in \mathbf{Q}(\beta)$, let T_x consist of all other algebraic integers which agree with x after the decimal point. The difference of any two elements of T_x is a series in positive powers of β. These act as contracting linear maps in the hyperplane S, so the projection of T_x to S has bounded diameter, independent of x. Each T_x has at least one representative in the slab $S \times [0,1]$. It follows that the closures K_x of the projections of the T_x to S overlap with bounded multiplicity.

Note that $\beta^{-1}T_x$ is a finite disjoint union of T_y. Therefore, one can express $\beta^{-1}K_x$ as a finite union, probably not disjoint, of K_y's (a subdivision rule). (This is a crucial point where it is important that β be a unit.) If the dimension of S is 2, that is, β is a cubic number, then multiplication by β is a genuine similarity of S. Otherwise, β^{-1} stretches the shape of K_y differentially in different directions.

We claim that there are only a finite number of the K_y's, up to translation in S. In order to see this claim clearly, it is a good time to introduce the language of finite state automata. A *finite state automaton* or *finite state machine* M over an alphabet A is a finite set S_M, (the set of states of M), a map $A \times S_M \to S_M$ (the state transition map for M), together with a distinguished element $I \in S_M$ (the initial state), and a distinguished subset $\mathrm{OK} \subset S_M$ (the accepting states). It is often convenient to visualize M as a labeled directed graph, with a bit of extra structure: I and OK. M operates by starting in its initial state; as it is fed elements of its alphabet one-by-one, it goes to the state indicated by its state transition map.

Given a word W in the alphabet A, W is *accepted* by M if when you start at I and go along the directions given by W, you end up in OK. The set of words $L(M)$ accepted by M is called the *language of* A. $L(M)$ is *prefix-closed* if every prefix of a word in $L(M)$ is also in $L(M)$, or in other words, if every (accessible) non-accept state has arrows only to another non-accept state. In such a case, we may as well collapse $S_M - A$ into a single fail state F, with all arrows leading back to itself. It is convenient, in drawing a picture for such an M, to omit the fail state and all roads leading to it. Whenever a word W gives you directions where there is no corresponding arrow, you immediately fail with no chance for reinstatement.

The definition for acceptance of an infinite word is not so clear in general, but if $L(M)$ is prefix-closed, there is an obvious definition: an infinite word is accepted if and only if each finite prefix is accepted.

PROPOSITION 9.5. PERIODIC CARRY FSA. *The set of weakly proper base β representations is the set accepted by a finite state machine M_β, if and only if* carry(β) *is eventually periodic*

PROOF: (See figure 9.4, proper FSA.) If $c = \mathrm{carry}(\beta)$ is eventually periodic, so that $c_{k+p} = c_k$ for all $k > q$, then let S_M be integers $0, \ldots, p+q-1$ together with a fail state F. The initial state is 0. From state $i < p+q-1$, $c_{i+1} \to i+1$, while all arrows with labels less than c_{i+1} lead to state 0, and all arrows whose labels are greater fail. From state $p+q-1$, c_{p+q} leads to state q, while all arrows with lower labels lead back to 0 and all arrows with greater labels fail.

Conversely, if there is an FSA M which recognizes all weakly proper base β representations, then we can reconstruct carry(β) from M by choosing, at each stage, the greatest

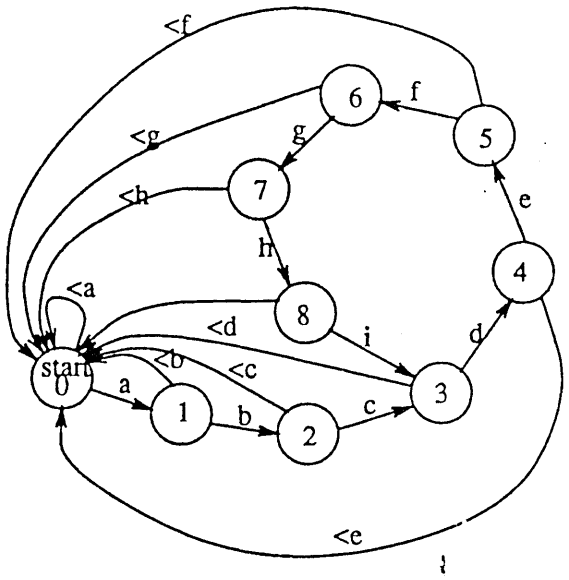

Figure 9.4. proper FSA. *This diagram illustrates a finite state machine for recognizing when a base β representation is proper, where* carry$(\beta) = .abc(defghi)$. *Each arrow back to 0 stands for a collection of arrows, one for each integer less than the indicated amount. Arrows not indicated lead to the fail state (not shown).*

digit accepted by M. Since M has only a finite number of states, the choices must eventually repeat.

9.5, periodic carry FSA

Given the right part r of x_β (after the decimal point), then the question of which left halves l satisfy that lr is a weakly-proper base β expansion depends only on the state which M is in after reading l. Let $F(x_\beta)$ be the set of states after which M accepts r: then $F(x_\beta)$ determines the shape of K_x, so there are only finitely many possibilities.

It does not quite follow that the K_x determine a tiling of the S, for they could in principle have substantial overlap. In fact, we have not given a definition of a tiling, let us do it: A *shingling* of a locally compact space X is a covering by a countable collection of compact sets (shingles) K_i, each equal to the closure of its interior, such that any compact subset $L \subset X$ only intersects finitely many K_i. A *tiling* of X is a shingling such that the intersection of the interiors of any two shingles is empty.

Note that the definition does not impose other topological restrictions on the shingles or tiles: they need not be connected, or locally connected, or simply-connected.

However, in many cases of this construction, the shinglings are tilings, and the tiles are disks.

We shall see later that every self-similar shingling is closely related to a self-similar tiling.

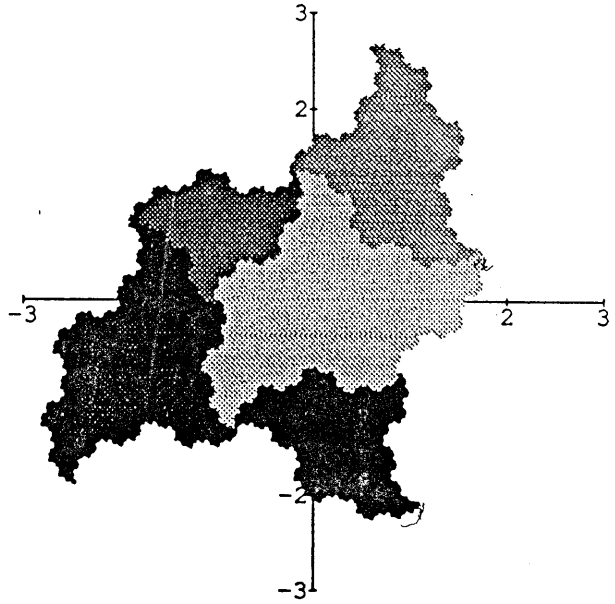

Figure 9.6. Pisot tiling of plane. *This tiling of a portion of the plain was obtained as the 'Galois dual' of the base α tiling of* $\mathbf{R}$, *where* $\alpha^3 = \alpha + 1, \alpha \approx 1.324717957244746$, *whose carry sequence is* .(10000). *This picture shows the projection to* $\mathbf{C}$ *of algebraic integers x in* $\mathbf{Q}(\alpha)$ *such that x_α has at most* 4 *nonzero digits to the right of the decimal point, (.0,.1,.01,.001,.0001) and at most about* 30 *to the left. They are shaded according to the portion to the right of the decimal point. The self-similarity of the tiling has contraction factor equal to one of the Galois conjugates of α, $\alpha_1 = -.6623589786\cdots - .56227951206\ldots i$, or reciprocally, expansion factor $-0.877438833\cdots + 0.7448617666\ldots i$. Compare this figure to 9.7, interpreting the five tiles as states of the FSA.*

Now we will generalize to complex numbers. If β is a complex number of modulus > 1, how can we define a base-β expansion for $\mathbf{C}$? First choose a finite set $D = \{d_1, d_2, \ldots, d_n\}$ of 'digits', with $0 \in D$. These could be $\{0, 1, \ldots, n\}$, or any other set of complex numbers.

Consider (β, D)-solitaire, defined as follows. At the beginning you are given a complex number z. (This is like a shuffled deck of cards.) You can subtract any element of D from z; then it is multiplied by β to get the next z. If z ever grows large enough, from then on, no matter what you do, it will grow larger, since multiplication by β dominates subtraction of elements of D for large z. In this event, you lose.

Let W be the set of initial z for which there exists a sequence of moves which do not lose. It is easy to see that W consists of all sums of series $\sum_{i=0}^{\infty} d_i \beta^{-i}$, $d_i \in D$. W is compact, contains 0, and satisfies the equation $W = D + \beta^{-1} W$: this is a characterization of W.

If D is too small, then W will be small. However, for any β there exist sets D such that W contains a neighborhood of the origin. We can, for instance, choose an arbitrary

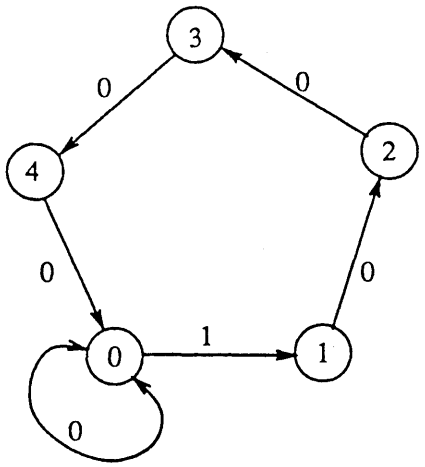

Figure 9.7. Pisot FSA. *The FSA which recognizes proper base-α representations, where α is the same as in figure 9.6.*

neighborhood U of the origin, and then make sure that D is large enough that $U + D$ contains βU. This guarantees that $U \subset W$, since we can move back into U after each move.

If W does contain a neighborhood of the origin, then every complex number admits a (probably non-unique) base (β, D)-representation, *i.e.*,, an expression $\sum_{i=i_0}^{\infty} d_i \beta^i$.

How can we select a preferred representation? First choose an ordering of D. With this choice, we can add another element of skill to (β, D) solitaire: now the object is to avoid shooting to infinity, while selecting the greatest possible digit at each stage (given previous choices.) In other words, the preferred or proper representation of a complex number z is the one which is greatest in lexicographical order. A representation is weakly proper if, for every finite initial segment, there is an extension which is preferred.

Note that this definition agrees with the previous definition in the case $\beta > 1$ is a real number, and $D = \{0, 1, n\}$ where n is the greatest integer less than β, with the natural linear ordering.

PROPOSITION 9.8. SOLITAIRE FSA. *Suppose that $\beta \in \mathbf{C}$ is an algebraic integer such that all its Galois conjugates except β and $\bar{\beta}$ are inside the unit circle in $\mathbf{C}$. If D is an ordered set of algebraic integers in $\mathbf{Q}(\beta)$, then there exists a finite state machine $M(\beta, D)$ which will recognize whether a sequence of $m = \{m_i\}$ gives a weakly preferred representation for some element $z \in W$.*

Remarks: The element z in question is of course $\sum_i d_i \beta^{-i}$.

We have not assumed that W contains a neighborhood of 0, even though that is the case of interest.

It is important that we are dealing here with series which begin with the β^0 term. If we allowed positive powers, there might not be any lexicographically latest element representing a complex number z.

This proposition is closely related to 11.9, hyperbolic automatic (Cannon).

PROOF: Consider any other sequence of moves $n = \{n_i\}$, and compare the result of n with that of m, starting at the same point z.

The difference of the two trajectories is the same as if we used the set of differences $D - D$ for digits, and started at 0, applying moves $n_i - m_i$.

Assuming that m-moves remain bounded, then the n-moves remain bounded if and only if 0 stays bounded using moves $n_i - m_i$. Let's now use the fact that we are in an algebraic number field $\mathbf{Q}(\beta)$, and consider the multi-dimensional picture $V = \mathbf{R} \otimes \mathbf{Q}(\beta)$. The difference moves begin at 0 and the moves are always algebraic integers, so it always remains an algebraic integer. Multiplication by β has a 2-dimensional invariant expanding subspace U, which we can identify with $\mathbf{C}$, and a complementary invariant contracting subspace S (by the hypothesis on β.) No matter what sequence of moves are applied from $D - D$, the point always remains in a bounded neighborhood of U.

On the other hand, if n is a competitor with m for representing z, then the point must also remain in a bounded neighborhood of S, that is, bounded in the complex plane. Therefore, in making comparisons with m, we can restrict to a compact subset of V, in which there is only a finite set A of algebraic integers.

The states of our machine M will consist of subsets of A: after a sequence of m-moves, the state will be the subset A_i of positions which are attainable by sequences of moves n lexicographically greater than m such that the difference sequence is always (until this moment) in A. The subset A_{i+1} is clearly determined by the subset A_i, together with the move m_{i+1}. The initial state is $\{\}$. Every set A_i which contains 0 is a fail state. There may be other fail states in addition: in general, define $B \subset A$ to be a fail state if $W \subset (B) + W$ (in $\mathbf{C}$).

9.8, solitaire FSA

Let us now again suppose that W contains a neighborhood of the origin. There is a sequence of tilings T_k of W, where the tiles are labeled by the initial k terms of weakly preferred sequences of moves, and a tile consists of the complex numbers represented by all weakly preferred sequences of moves beginning with those k terms. Each of these sets has non-empty interior, and its shape up to similarity depends only on the state of the machine $M(\beta, D)$ after reading its label. The shape up to translation depends only on the state, together with k.

If we expand T_k, multiplying by β^k, the shapes of tiles only depend on a state of $M(\beta, D)$. Each tile has a rule for subdivision, given by the state transition rules for $M(\beta, D)$.

To obtain a self-similar tiling, choose a k and a tile which occurs in the interior of the kth subdivision of itself. Expand this kth subdivision by β^k, and translate it so that the chosen tile coincides with the original. Repeat this subdivision/expansion/translation process indefinitely, to obtain a self-similar tiling of the plane with expansion constant β^k.

It may not be possible to obtain a self-similar tiling with expansion constant β with this particular set of tiles. However, if we choose the linear ordering of the digits to make 0 *greatest*, then the 0-arrow from the initial state (the empty set of competitors) leads back to the initial state. In this case, the sequence of tilings $\beta^k T_k$ agree where they overlap, so their union is a self-similar tiling of $\mathbf{C}$.

There is another picture associated with these tilings, which helps put it in context *vis. a vis.* dynamical systems. Let us consider the case first that β is an algebraic unit,

so that multiplication by β acts as an automorphism of the lattice of algebraic integers $\Lambda \subset V$. Therefore β induces a diffeomorphism ϕ_β of the quotient space, V/Λ, which is a torus. Since no eigenvalues of the linear transformation are on the unit circle this is what is called an Anosov diffeomorphism of the torus. The invariant subspaces U and S together with the planes parallel to them map into the torus to define two foliations, F^u of dimension 2 and F^s of dimension $d-2$, of the d-torus. These foliations are invariant by ϕ_β.

The theory of hyperbolic dynamical systems tells us that in this situation, there is a Markov partition for ϕ_β, that is, a finite cover by closed sets R_j, each of which is a product in local coordinates of a set of leaves of F^s and a set of leaves of F^u, such that the R_j have disjoint interiors, and when the interior of $S_i = \phi_\beta(R_i)$ intersects R_j, it stretches clear across R_j in the F^u direction and squeezes inside R_j in the F^s direction. Another way to say this is that the intersections of the sets R_j with a generic leaf of F^u or a generic leaf of F^s defines a tiling of the leaf; the Markov property says that ϕ_β acting on an unstable leaf maps each tile to a union of tiles, and on a stable leaf maps it to a subtile. There are only a finite number of tile types, since the tile type is determined by R_i. If there is a 'generic' unstable leaf l which is mapped to itself, its induced tiling is a self-similar tiling of the plane.

More generally, if β is an algebraic integer but not necessarily a unit, there is an associated map ϕ_β of the torus $T = V/\Lambda$ to itself, as before, but it may be n-to-1. However, we can form the inverse limit of the sequence of maps

$$\ldots T \to T \to T \to T$$

to obtain a compact space T_β (a Cantor set bundle over the torus), on which the inverse limit map $\bar{\phi}_\beta$ acts as a homeomorphism. The action of ϕ_β is still hyperbolic, and T_β has two foliations, F^u and F^s which are invariant. The leaves of F^u are homeomorphic to the complex planes, but the leaves of F^s are homeomorphic to $\mathbf{R}^{d-2} \times C$, where C is a Cantor set. Again, the general theory of dynamical systems implies that a Markov partition for ϕ_β exists. It yields almost self-similar tilings of $\mathbf{C}$, and with some added care, actual self-similar tilings.

§10. Characterization of expansion constants

We have been discussing and defining self-similar tilings by example and by context, but now we give a more formal definition.

A tiling T of the complex plane $\mathbf{C}$ is *self-similar* with *expansion constant* $\lambda \in \mathbf{C}$ if
(a): The tiles of T can be divided into a finite number of distinct 'types', such that tiles in a given type differ only by translations of the plane.
(b): When T is mapped by multiplication by λ, the image of each tile is a union of tiles.
(c): The pattern (relative positions, shapes, and types) of subdivision of the image of any tile under multiplication by λ depends only on the type of the tile.
(d): The tiling is quasihomogeneous, that is, for any $r > 0$ there is an $R > 0$ such that for every disk D of radius r in $\mathbf{C}$ and every disk E of radius R, an isomorphic copy of D (including types) can be found within E.

Some of these points bear discussion. It would be interesting to relax condition (a), to say that tiles of the same type are congruent, but not necessarily by a translation of the plane. I would conjecture that no more examples are obtained by this relaxation. The types of tiles are not to be regarded as part of the structure of the tiling: they are a convenience for dealing with the rules of subdivision. When two tiles have the same type, it implies that they are congruent, that their subdivisions are congruent, and that all subsequent subdivisions are also congruent.

Condition (d) is imposed to avoid problems one encounters in certain cases that are not of real interest anyway. For instance, consider a tiling of the plane by squares of two sizes, say 1 and π, with tiles of size 1 to the left of the y-axis and size π to the right. This can be constructed so that expansion by a factor of 2 takes each tile to a union of 4 tiles. It satisfies (a), (b), (c) but not (d).

There are also many examples of tilings where a tiling is not strictly self-similar, but where there is a cycle of tilings $T_0, T_2, \ldots, T_{p-1}$ such that $T_{i+1 mod p}$ is a subdivision of the expansion of T_i, using rules depending only on tile types, as above. Such a tiling will be called *periodically self-similar*. There are also interesting still weaker conditions which we will not address now.

In this section we will prove:

THEOREM 10.1. LARGEST INTEGERS EXPAND TILINGS. *A complex number λ of modulus bigger than 1 is an expansion constant for some self-similar tiling if and only if λ is an algebraic integer which is strictly larger than all its Galois conjugates other than its complex conjugate.*

Remarks: This theorem gives examples much more general than in the preceding section.

The corresponding condition for periodically self-similar tilings, which we will not prove, is that all Galois conjugates of λ have modulus less than or equal to that of λ, and that those of the same modulus have a ratio with λ or with $\bar{\lambda}$ which is a root of unity.

This theorem and its proof generalizes fairly easily to arbitrary dimensions by talking about linear transformations λ and their Galois conjugates. The dimension 1 case is essentially the Perron-Frobenius theorem and its 'converse' of Doug Lind ([**Lind**]).

The rate of growth of area is $\lambda\bar{\lambda}$, which is a real number larger than all its Galois conjugates: this fact is actually an easy consequence of the Perron-Frobenius theorem. It is not enough to guarantee a self-simialr tiling, for there are examples of algebraic integers such that $\lambda\bar{\lambda}$ is larger than its Galois conjugates, but λ has Galois conjugates bigger than itself. In such a case, the Galois group is necessarily smaller than the symmetric group.

The minimum number of tiles for a selfsimilar tiling of expansion constant λ is at least the maximum of the degree of λ and the degree of $\lambda\bar{\lambda}$. This is not sharp lower bound, however,

PROOF: The easier direction is the 'only if' direction, so we will do that first.

Let T be a self-similar tiling with expansion constant λ. The proof can be thought of in terms of establishing a system of governance and a system of roads for the countryside of T.

We will first choose a *capital* (or capital) for each tile, in such a way that the capital

of any tile maps to the capital of another tile under multiplication by λ, and so that the position of the capital relative to a tile depends only on its type.

To do this, we can graphically represent the rule for subdivision of types of tiles as a directed graph Γ. The nodes of Γ are labeled by the types of tiles, if a given type x occurs k times in the subdivision of another type y, k edges of Γ lead from y to x, with each edge corresponding to a relative position of one of the x-tiles within the expansion of y.

For each node of Γ, choose one distinguished outgoing edge.

We impose the condition that the capital of any tile maps under expansion to the capital of the tile pointed to by its distinguished outgoing edge.

This determines the capital $c(t)$ uniquely for every tile t, since in the sequence of subdivisions of a tile, the subtiles necessarily shrink to points.

Now define a set D to consist of all differences of capital cities of townships and counties,

$$D = \{c(s) - \lambda c(t)\}$$

where s is a tile contained in λt. These differences are determined by edges of the graph Γ, so there are only a finite number. Labeling the edges of Γ by the appropriate elements of D, we almost have a finite state machine: we make such a machine M by adding a special initial state I, a special fail state F, and adjoining to the alphabet D special symbols begin_t, where t ranges over the tiles which contain 0. (If 0 is in only one tile, this is not necessary. In the decimal system, $+$ and $-$ play an analogous role to begin_t, for the tiling of $\mathbf{R}$ by intervals between integers, $\lambda = 10$.)

As usual, we use the convention that if there is no outgoing arrow with label d from a node t, then d leads to the fail state, and that all non-fail states are accept states.

It is worth observing that the tiling can easily be reconstructed from M. In fact, M determines a base λ-system for $\mathbf{C}$, where

$$\text{begin}_t\, z_k \ldots z_0 . z_{-1} z_{-2} \ldots$$

is proper if and only if the sequence of digits is accepted by M, and it represents the complex number

$$z = \sum_i z_k \lambda^k.$$

The tiles are labeled by the 'whole' part of this expansion (to the left of the decimal point) and consist of all complex numbers z sharing the whole portion.

So far we have developed enough structure to connect the capitals in a hierarchical grouping, but there are no provisions for tourism or commerce. Next we need to build enough roads to connect the capitals of neighboring tiles in a reasonably efficient manner. To this end, we would like to find a finite set R of 'enough' differences between capitals, enough so that you can go from any capital to any other along roads of type R quasi-efficiently.

More formally, we stipulate that R is sufficiently large that there exists a constant K such that for any two capitals c_α and c_ω in $\mathbf{C}$, there is a finite sequence $\{c_\alpha = c_0, \ldots, c_n = c_\omega\}$,

where $c_{i+1} - c_i \in R$ and $n < K|c_\omega - c_\alpha|$. (In this definition, capitals of distinct tiles are considered identical if they are in the same place.) One way to guarantee this is to let R consist of all differences of capitals of tiles which have distance no greater than 3 times the maximum diameter X of a tile. It is obvious that one can get around quasi-efficiently with such roads. It follows from the quasi-homogeneity property (c) in the definition of a self-similar tiling that R is finite.

Consider now the effect of multiplication by λ. For each road $r_i \in R$, λr_i is a difference of two capitals that are somewhat farther apart, so it can be expressed as $\lambda r_i = \sum_j h_{ij} r_j$, where the h_{ij} are integers. In other words, the vector $(r_1, \ldots, r_m)$ is an eigenvector of the 'highway rewriting matrix' $H = (h_{ij})$, with eigenvalue λ.

It follows at once that λ is an algebraic integer: it satisfies the characteristic equation for H.

There are actually many different routes between any two capitals. We now eliminate this ambiguity. Let J be the additive subgroup of $\mathbf{C}$ generated by R. Since J is a finitely generated torsion free abelian group, it is isomorphic to $\mathbf{Z}^l$ for some l. Let $j_1, \ldots, j_l$ be generators for J. The j_i are not necessarily in R. The difference between any two capitals $c_1 - c_2$ can be expressed in a unique way as an integer linear combination of the j_i. The sum $N(c_1 - c_2)$ of the absolute values of the coefficients is less than some constant times the minimum number length of a chain of roads between capitals, so it is less than a constant times the $|c_1 - c_2|$. On the other hand, since there is an upper bound to the length of any j_i, $N(c_1 - c_2)$ is also greater than some positive constant times the $|c_1 - c_2|$.

Multiplication by λ induces an endomorphism of J, which we also denote H. If r is any road, it follows that

$$\lim_{k->\infty} \frac{1}{k} \log(N(H^k(r))) = |\lambda|.$$

In other words, the geometric growth rate of the images of r under the endomorphism H of J is λ. But the geometric growth rate of any vector r under iteration of a linear transformation is the largest modulus of a characteristic root of the linear transformation, restricted to the invariant subspace generated by r. Therefore λ is at least as great as any of its Galois conjugates.

To complete the 'only if' portion of the proof, it remains to show that λ is strictly larger than all other Galois conjugates except its complex conjugate. We will see this by looking at our multi-dimensional picture, $V = J \otimes \mathbf{R}$, a little more. Suppose that μ is any other characterisic root of H such that $|\mu| = |\lambda|$. There is then a linear map $p_\mu : V \to \mathbf{C}$ which conjugates the action of H to multiplication by μ (in fact, p_μ is a μ-eigenvector of the dual linear map acting on $\mathbf{C} \otimes V^*$.) Let p denote the original linear projection of V to $\mathbf{C}$.

Define $\mathcal{D}_i \subset \mathbf{C} \times \mathbf{C}$ to be the image of $H^{-i}(\mathcal{C})$ by $p \times p_\mu$, and let $\mathcal{D}$ be the closure of the union of $\mathcal{D}_i$.

Claim: The projection of $\mathcal{D}$ to the first factor is a homeomorphism. $\mathcal{D}$ is the graph of a Lipschitz map f : *complexes* $\to \mathbf{C}$ conjugating multiplication by λ to multiplication by μ.

The claim is easy to see. In fact, the original set $\mathcal{D}_0$ clearly satisfies a global Lipschitz condition, that is, there is a constant K such that for any $(x_1, y_1) \in \mathcal{D}$ and $(x_2, y_2) \in \mathcal{D}_0$, $|y_2 - y_1| \leq K|x_2 - x_1|$ (because of the quasi-efficiency of the system of roads). But $\mathcal{D}_i$ is the image of $\mathcal{D}_0$ under the map (λ_{-i}, μ^{-i}) which multiplies distances exactly by $|\lambda^{-i}|$.

Therefore the Lipschitz condition is uniform in $\mathcal{D}$.

A Lipschitz map f is differentiable almost everywhere. Let z be a point where it is differentiable. If we expand a neighborhood of z by a high power n of (λ, μ), then the image point has a neighborhood of a given radius r where the graph of f is very close to being linear. By the quasi-homogeneity of the tiling, it follows that there is some point z' in a neighborhood of radius $R(r)$ of the origin. (This follows because the portion of $\mathcal{D}$ above any tile is determined by the type of the tile.) Fixing r and taking the limit as $n \to \infty$, we obtain a point which has a neighborhood where f is exactly linear. If r is large enough, then the disk of radius 1 about the origin is contained in this disk of radius r. Therefore, f is linear. It follows that $\mu = \lambda$ or $\mu = \bar{\lambda}$.

This completes the only if portion of the proof.

The second half of the proof will be, given an algebraic integer λ such that all its Galois conjugates except λ and $\bar{\lambda}$ are smaller, to construct a selfsimilar tiling.

The construction will naturally make use of the vector space $V = \mathbf{R} \otimes \mathbf{Q}(\lambda)$.

The aim is to construct a subset $\mathcal{C} \subset \Lambda \subset V$, where Λ denotes the lattice of algebraic integers in $\mathbf{Q}(\lambda)$, such that $\mathcal{C}$ is self-similar in some appropriate sense, and so that it projects to a discrete, quasihomogeneous set in $\mathbf{C}$.

We may as well start with $0 \in \mathcal{C}$. Now pick a few more elements of Λ to be in $\mathcal{C}$, enough so that 0 is in the convex hull of the projection of the given points to $\mathbf{C}$. Define this set to be $\mathcal{C}_0$.

(Sketchy at the moment.)

Iteratively expand and interpolate.... Use the Delaunay triangulation to decide when to interpolate ... Make a deterministic rule, depending only on the shapes of the DeLaunay triangles, together with the time since creation — that is, wait a while before subdividing, then subdivide thoroughly.... Make a hierarchical structure: each new vertex is associated with the vertices of the previous triangulation nearest The resulting tiles have good quality if one waits a long time before subdividing, but there may be very many of them. However, this picture is not yet quasi-homogeneous.... modify the construction, by choosing a cyclic ordering of the vertex types that occur (with some bounds on shapes and eccentricities of Delaunay triangles): and put a tiny copy of the successor vertex type, as the first step in making the choice for each vertex type.

10.1, Largest integers expand tilings

§11. Automatic groups

Like many things in mathematics, groups can be difficult to get a handle on. In fact, a celebrated result of Novikov and Boone says that there is no general algorithm, given two presentations for groups, to tell whether or not they are isomorphic: it is not even possible to tell whether a presentation describes the trivial group. Furthermore, there are particular presentations for which, given two words in the generators, it is not possible to tell whether or not they represent equal elements in the group — or equivalently, given a single word, it is not possible to tell whether or not it equals 1.

It is worth emphasizing that the difficulty is not in finding an algorithm which will answer 'yes, they are equal' if they (the groups, or the words) are equal. Such algorithms,

in fact, are easy to construct, although they tend to be stupid and incredibly slow: the idea is simply to try all possibilities. The difficulty is in finding algorithms which will answer 'no they are not equal' if they are not.

Despite the fact that intractable groups exist, we do not need to be discouraged about finding techniques which might apply to the many particular groups which we would like to understand better.

The theory of automatic groups is one attempt to delineate a reasonably large class of groups, including many that arise in real mathematical contexts, where it is indeed possible (but not necessarily easy) to 'see' what they look like and to analyze them algorithmically by computer.

Here is the formal definition of an automatic structure for a group; we will illustrate it by examples and interpret it more geometrically later:

An *automatic structure* for a group G is
(a). a set of generators $\mathcal{G}$ for G,
(b). a set of words R accepted by some finite state automaton WA (the word acceptor) with alphabet $\mathcal{G}$, containing at least one word representing each element of G, such that
(c). for each element $g \in \mathcal{G}' = Gen \cup \{\$\}$, there exists a finite state automaton C_g (the g-comparator) with alphabet $\mathcal{G}' \times \mathcal{G}'$. Given a word $w = (u_1, v_1)(u_2, v_2)\dots(u_n, v_n)$, let u be the word $u_1u_2\dots u_n$, and let $v = v_1v_2\dots v_n$. Then C_g accepts w if and only if u and v are \$-free prefixes u' and v' followed by (possibly empty) strings of the pad symbol \$, where u' and v' are each accepted by WA and $v' = u'g$ in G.

The word acceptor automaton (b) should be thought of as picking out canonical forms for group elements, although this canonical form need not be unique. The comparator automata of (c) can be thought of as knitting the canonical forms together, to construct the the group.

Note that, according to the definition, the automatic structure is defined by the set of generators together with the set of words R: a word acceptor WA and comparators described in (c) must exist, but they don't have to be produced to define the structure.

Part (c) of the definition in particular may seem technical and opaque now, but we will soon deduce an equivalent condition which is more intuitively comprehensible. Before doing that, however, let's look at least at a trivial example using these definitions.

First consider $\mathbf{Z}$, with generators $a = 1$ and $A = -1$. The word acceptor has three non-fail states

Since only one word is accepted by WA for each element of G, the comparator C_1 just recognizes whether two word are equal (and accepted by M) The comparator C_a (11.2) has four non-fail states, of which only one is an accept state.

What sorts of canonical forms (specified by WA) can there be? Many of you may be familiar with them in another guise: the set of words (the *language*) accepted by a finite state automaton is what is called a *regular language* or a regular set. Regular sets are commonly used in many word-processing applications on computers. Typically you specify a regular set by a *regular expression* or pattern, and the program constructs a finite state automaton, sets it running on your file, finds matches (that is, strings fitting your pattern), and prints them out, makes substitutions, or whatever you asked it to do.

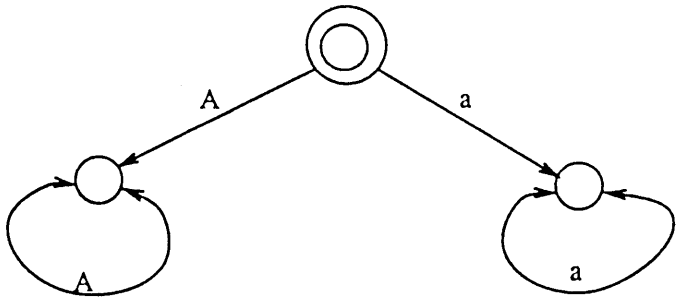

Figure 11.1. Z word acceptor. *A word acceptor for* Z*, with presentation* $\langle a|\rangle$*. We implicitly assume that the generating set is closed under inversion, and that change of case denotes inverse, unless otherwise noted.*

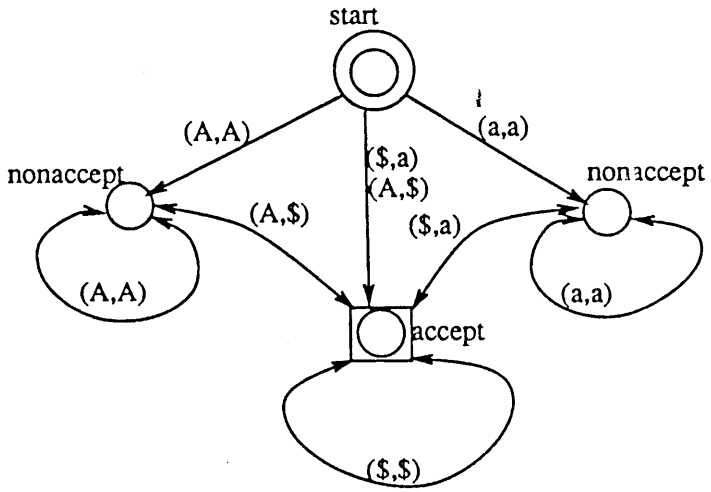

Figure 11.2. Z comparator. *This is the comparator* C_a *for the group* Z*, generated by a. It has only one accept state, and three other states from which it is possible to reach the accept state. Any arrow not shown leads to the fail state, not shown, from which it is impossible to escape.*

A good example is the Unix utility egrep. The word acceptor for Z, for instance, could be specified by the regular expression

```
a*|A*
```

where the symbol * denotes zero or more repetitions of the preceding object, and the symbol | means 'or'. The command

```
egrep '^a*|A*$'
```

prints out all lines of its inputs which are accepted by WA. The symbol ^ here denotes the beginning of a line, $ denotes the end of a line, and parantheses are used for grouping.

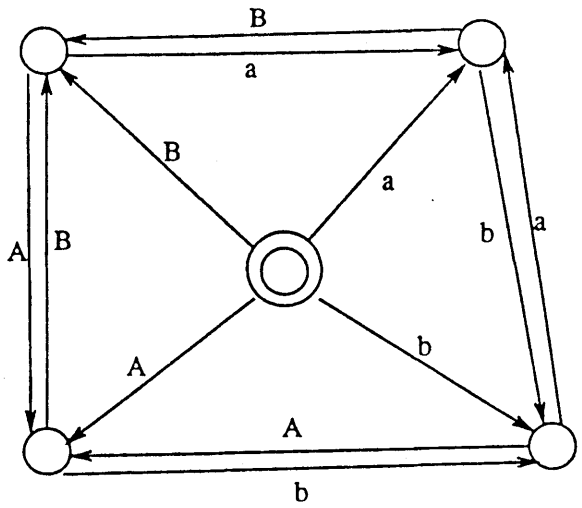

Figure 11.3. Free word acceptor. *A word acceptor for reduced words in the free group* $\langle ab|\rangle$.

(The quotes ' protect all the special symbols from being interpreted by the shell (command parser), before egrep gets them).

It is also easy to construct an automatic structure by inspection for the group on n generators. For instance, a word acceptor which accepts only words in reduced form for the free group $\langle ab|\rangle$ is illustrated in 11.3. The corresponding egrep command is

```
egrep '^(b+|B+)?((a+|A+)(b+|B+))*(a+|A+)?$' .
```

Here some more notation has been introduced. Parentheses '()' are used for grouping. The operator ? means zero or one occurrence: (*expression*)? is equivalent to ((*expression*|)). The operator + means one or more occurrences: (*expression*)+ is equivalent to (*expression*)(*expression*)*. The egrep command asks for a word whose main part consists of repeated strings of a's or A's followed by a strings of b's or B's, and is matched by ((a+|A+)(b+|B+))*. However, the beginning and ending might be in a different phase, hence the extra stuff enclosing it.

The reader might enjoy constructing the egrep expression for reduced words in the free group on three generators.

Of course, this use of egrep is not the use for which it was designed, and the regular expressions for a group tend to be a bit long-winded. Nonetheless, the efficiency and the success of egrep and other related code is an inspiration and a guide to what we may be able to accomplish with groups.

Geometrically, a word in the generators of a group G is equivalent to a simplicial path in the group graph $\Gamma(G)$ (see §3, Group graphs) starting at the base point. The set of words accepted by a word acceptor automaton thus defines a family of paths in the graph of the group beginning at the base point, at least ending at each vertex.

The fact that different paths have different domains is an inconvenience here. If $\gamma : A \to X$ is a path, where A is an interval, let $\bar{\gamma} : \mathbf{R} \to X$ be the extension to all of $\mathbf{R}$ which is constant in the components of the complement of A.

We can define a *combing* of a metric space X to be a family F of paths in X beginning at the base point, including the trivial constant path, such

(a): there is a constant $K > 0$ such that for any $x \in X$ at least one path ends within a distance of K of x, and
(b): for any $L > 0$ there is an M such that whenever two paths end within a distance of L, then they are within distance M for all time.

A combing, in other words, is approximately a uniformly continuous right inverse to the map which takes a path to its endpoint, using the uniform metric on paths. However, the inverse need be defined only sketchily, and it need not be continous: its discontinuities are bounded (by K), however. If we didn't allow discontinuities, combings could exist only for contractible spaces. (We could then try to work with classifying spaces for groups, rather than graphs of groups, but we have no guarantee that the groups we will deal with have compact classifying spaces.)

PROPOSITION 11.4. AUTOMATIC COMBING. *A set of generators $\mathcal{G}$ for a group G together with a set of words R satisfying condition (b) gives an automatic structure for G if and only if the paths in $\Gamma(G)$ defined by R is a combing of $\Gamma(G)$.*

PROOF:

Suppose, first, that the set of paths defined by R is a combing of $\Gamma(G)$. Let WA be a finite state automaton which accepts words in R, with padding by \$ at the end permitted. Let M be a constant such that any two paths ending within distance 1 of each other (that is, on adjacent vertices in the group graph) remain within a distance M for all time. Define a finite state machine $Diff$ with alphabet $\mathcal{G}' \times \mathcal{G}'$, whose set of states is the set of group elements within a distance of M from the identity together with a fail state. On reading a pair of words (u, v) (combined, as in the previous discussion, to make a single word in $\mathcal{G}' \times \mathcal{G}'$) the state of $Diff$ at any time is Fail if either of the component words is not accepted by WA or if the words at some time have been at a distance greater than M from each other; otherwise, the state of $Diff$ after reading k symbols is the difference, $u_k^{-1} v_k$ where w_k denotes the length k prefix of a word w. The non-Fail transitions of $Diff$ on input (a, b) go from state g to state $a^{-1}gb$.

Comparator machines can be obtained from $Diff$ just by choice of the which states are accepted: the only accept state for C_g is g. This shows that if R defines a combing of $\Gamma(G)$, then R gives an automatic structure.

Suppose, conversely, that R determines an automatic structure for G. To show that R defines a combing, it will suffice to prove that any two accepted words ending within a distance of 1 from each other remain a bounded distance apart, since words whose ends are more distant than 1 can be joined by a chain of words ending 1 apart. Thus, we need to show that for each comparator C_g, the pairs of words accepted by C_g remain a bounded distance apart.

If there are states in C_g which never can lead to an accept state, no matter what the input, we may collapse all such states to a single fail state, without changing the set of

words accepted by C_g.

Once this is done, we claim that the word difference $u_k^{-1}v_k$ depends only on the state of C_g after reading (u_k, v_k), provided this state is not a fail state. Indeed, suppose that the state of C_g after reading another pair of words $(u_l', v_l'$ is the same as the state after reading (u_k, v_k), and that this state is not a fail state. Then there is some suffix (w_j, x_j) such that C_g accepts $(u_k w_j, v_k x_j)$ and therefore also $(u_l' w_j, v_l' x_j)$. By definition of a g-comparator,

$$(u_k w_j)^{-1}(v_k x_j) = g = (u_l' w_j)^{-1}(v_l' x_j)$$

, and therefore

$$u_k^{-1} v_k = w_j^{-1} g x_j = u_l'^{-1} v_l',$$

that is, the word differences are equal.

Since C_g has only finitely many states, the set of possible word of accepted pairs of words is finite, hence their distance is bounded.

Therefore, the set of paths defined by R gives a combing of $\Gamma(G)$.

11.4, automatic combing

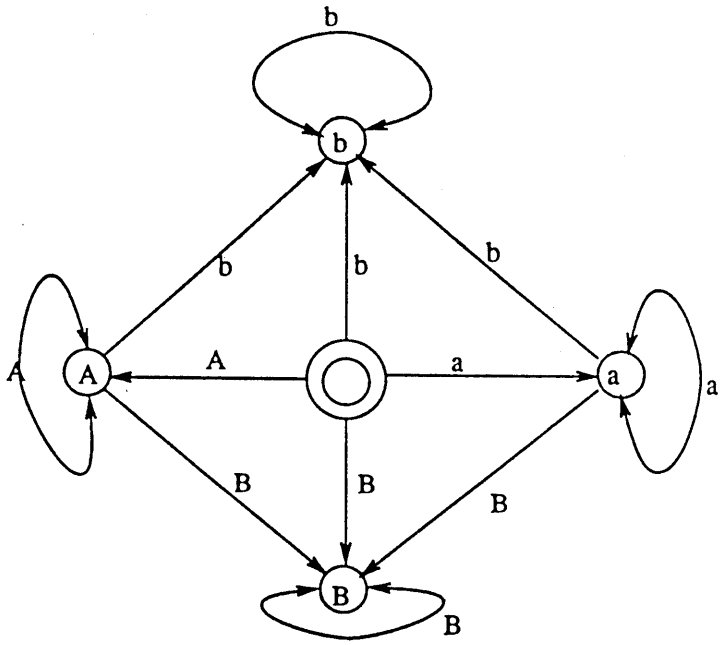

Figure 11.5. Abelian acceptor. *A word acceptor automaton for* $\mathbf{Z}^2$*, accepting words matching the regular expression* (a*|A*)(b*|B*)*. Note the resemblance to the acceptor for* $Z * Z$ *(11.3).*

As a simple illustration of this principle, let us construct an automatic structure for $\mathbf{Z}^2 = \langle a, b | abAB = 1 \rangle$. We can define the set R of accepted words to be those matched by the pattern

$$(\mathtt{a}*|\mathtt{A}*)(\mathtt{b}*|\mathtt{B}*).$$

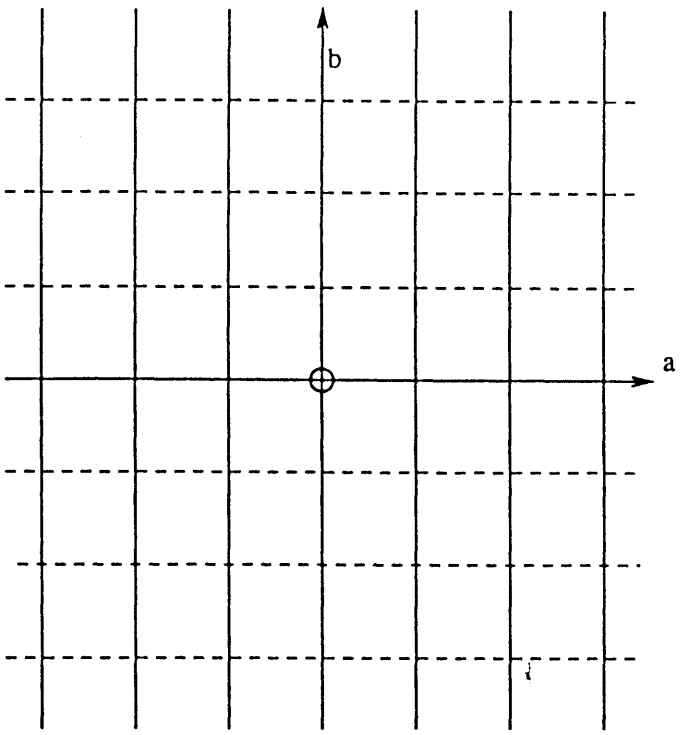

Figure 11.6. Abelian tree. *The word acceptor of 11.5 accepts reduced words which lie in the tree in the graph of the group* $\mathbf{Z} \times \mathbf{Z}$ *illustrated above.*

These are accepted by the simple finite state machine of figure 11.5. The words in R correspond to paths along the solid lines (horizontal, then vertical) of figure 11.6. Clearly these words form a combing, so R defines an automatic structure for $\mathbf{Z}^2$.

Often a major difficulty in handling finitely presented groups is to come up with concepts which are independent of the generating set. We are in reasonably good shape here:

PROPOSTION 11.7. AUTOMATIC INDEPENDENT OF GENERATORS. *If a group has an automatic structure with using one set of generators, then it has an automatic structure using any other.*

PROOF: This is quite easy. Suppose we have an automatic structure using generators $\mathcal{G}$, and that $\mathcal{G}_1$ is an alternate set of generators. For each $g \in \mathcal{G}$, choose a word w_g in $\mathcal{G}_1$ representing g. If R is the regular set of words for the original automatic structure, let R_1 be the set of words obtained by replacing each generator g by w_g. Clearly R_1 is recognized by a finite state machine WA_1: it can be constructed from WA by subdividing each edge labeled g by inserting new states so that it can be labeled by the elements of w_g.

Since R defined a combing, clearly R_1 also defines a combing (even though the graph $\Gamma(G)$ has changed, and the metric has changed, the metric induced on G has changed only by a bounded factor.)

11.7

To think about the geometry of a group in a way that is independent of choice of

generators (or other additional structure), one should try to understand the *quasi-geometry* of the group. A choice of generators for G defines a metric on G, the word metric, where the distance between two group elements g and h is the minimum length of a path in $\Gamma(G)$ joining g to h, or equivalently, the minimum length of a word representing $g^{-1}h$. When the set of generators is changed, this metric changes by a map satisfying some global Lipschitz condition: the metric changes (up or down) by a bounded factor.

A *quasi-geodesic* in a metric space X is a path $\gamma : A \to X$ (where A an interval) which in the large has a percentage efficiency bounded away from 0: that is, there is a constant K such that for any two real numbers $t_1 < t_2$,

$$d(\gamma(t_1), \gamma(t_2)) > 1/K(t_2 - t_1) - K.$$

The paths in any combing of X are quasi-geodesics. The set of quasi-geodesics of X depends only on the quasi-geometry of X.

If Q is any compact, connected space with fundamental group G and if Q has a path metric, that is, a metric in which the distance between points is equal to the minimum length of a path joining them, then the universal cover $\tilde{Q}$ has an induced path metric. The set of preimages of the basepoint in $\tilde{Q}$ is canonically isomorphic to G, so an induced metric is defined on G. This induced metric is clearly in the same quasi-class as any word metrics on G.

One case of particular interest is that K is a manifold of negative curvature: for instance, a hyperbolic manifold. Then the quasi-geodesics in $\tilde{K}$ are particularly nice:

PROPOSITION 11.8. HYPERBOLIC QUASI-GEODESICS NEAR GEODESICS. *Let M be a compact manifold or orbifold of strictly negative curvature, possibly with convex boundary. There is an L such that any finite K-quasi-geodesic γ in $\tilde{M}$ lies in the L-neighborhood of the geodesic g joining its endpoints. If the domain of γ is infinite in either or both directions, there is a unique limiting geodesic g in $\tilde{M}$ within a bounded distance of γ and ending at any finite endpoint of γ.*

This is a widely useful fact, which was used, for instance, in Mostow's rigidity theorem and many other places. We will not go over the proof here: see [**Thurston1**] for a proof. It is in striking contrast to the situation in Euclidean space. For instance, in the plane, a logarithmic spiral is a quasi-geodesic: the distance from the geodesic between points along it is unbounded. Intuitively, in hyperbolic space, as you move away from a geodesic, distances increase exponentially. If you wander very far away from a geodesic and then come back, then you are forced to retrace your route closely enough that some segment of your path has a very low efficiency.

An orbifold is a generalization of a manifold. It contains the appropriate structure to describe the quotient space of the action of a discrete group action where some elements of finite order may have fixed points. This is really independent of the thrust of the discussion here, so we won't explain further: if you are not already familiar with it, it is inessential.

Negatively curved with convex boundary have the property that for any two points in the manifold, any homotopy class of arcs between them contains a unique geodesic.

In 2 and 3 dimensions, most closed manifolds (or orbifolds) have metrics of negative curvature with convex boundary.

Here is a key existence theorem, which yields many automatic structures:

THEOREM 11.9. HYPERBOLIC AUTOMATIC (CANNON). *If M is any compact negatively curved manifold or orbifold, possibly with convex boundary, and if $\mathcal{G}$ is any set of generators for the fundamental group of M, then the set $L_{\mathcal{G}}$ of shortest words in $\mathcal{G}$ representing a given element of $\pi_1(M)$ is a regular set, and it defines an automatic structure for $\pi_1(M)$.*

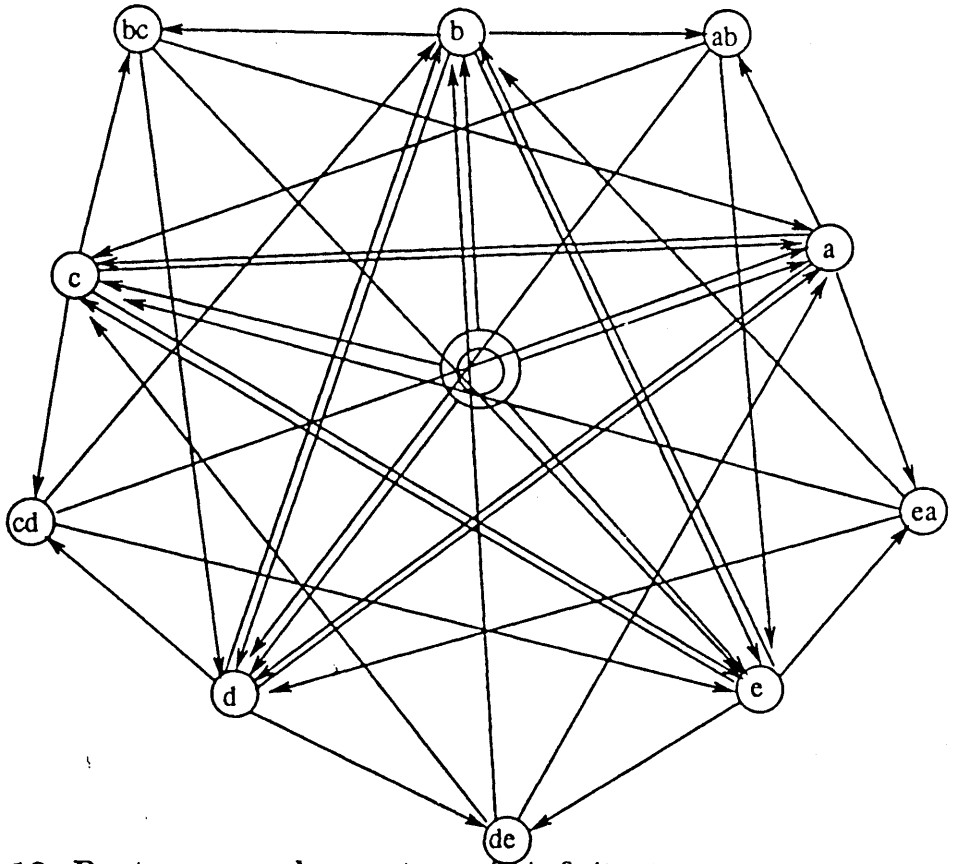

Figure 11.10. Pentagon word acceptor. *This finite state automaton accepts shortest words for the group of reflections of a right-angled pentagon in the hyperbolic plane,*

$$\langle a, b, c, d, e | aa, bb, cc, dd, ee, abab, bcbc, cdcd, dede, eaea \rangle .$$

The start state is the double circle in the middle. Each arrow leading into the state labeled with a single letter x is an x-arrow. Arrows leading into states labeled with two letters are either x arrows or y arrows; you can tell which by the condition that an arrow leading away from a state labeled x cannot be labeled x.

Remark This is closely related to 9.8, solitaire FSA, and also to 11.4, automatic combing.

PROOF: Let G be the fundamental group of a negatively curved compact orbifold with convex boundary, and let $\mathcal{G}$ be any set of generators. If w and v are any two shortest words in $\mathcal{G}$ representing elements of G within distance 1 of each other, then the paths they define in $\Gamma(G)$ are quasi-geodesics. Therefore there is some constant L such that for any such pair, the word differences $w_k^{-1}v_k$ have minimum word length less than L.

To construct a finite state automatan M which will recognize shortest geodesics, let the set of states S_M be the set of subsets of the ball of radius L, B_L, in $\Gamma(G)$, together with a fail state. The state s of M upon reading w_k is either the fail state, or it will be the set of all elements $w_k^{-1}v_k$ of G where w and v have remained within distance L of each other up to the current time k. If the $k+1$st generator g of w is in s, then the new state is the fail state. Otherwise, the new state is $g^{-1}s\mathcal{G} \cap B_L$.

This shows that the set of shortest words is a regular set. They form a combing, by , hence they define an automatic structure. 11.9

COROLLARY 11.11. HYPERBOLIC AUTOMATIC TREE. *The fundamental group of a compact, negatively curved manifold or orbifold with convex boundary admits an automatic structure such that the set R of accepted words is prefix-closed and represents each element of the group exactly once. In other words, R defines the set of simple paths in a spanning tree for the graph of the group.*

PROOF: Let R be the set of shortest words which is lexicographically least among all words representing the given element. A slight modification of the machine described in the proof above will select elements of R. 11.11

Gromov has developed a more abstract notion of a 'hyperbolic group'. There are many equivalent characterizations, but one characterization is that a hyperbolic group is a group satisfying the conclusion of Proposition 11.8, hyperbolic quasi-geodesics near geodesics. Such groups are therefore automatic.

The proof of 11.9 is constructive, but an algorithm which literally follows the proof would be extremely impractical. In the first place, it is not easy to get good constants for proposition 11.8, hyperbolic quasi-geodesics near geodesics. From it, one gets some constant L. In a hyperbolic group, the number of elements of word length less than L generally grows exponentially with L, so the size of B_L may be quite large. (There are trivial exceptions, from 0 and 1 dimensions: for instance, $\mathbf{Z}$ is a hyperbolic group.) Finally, the set S_M has cardinality $2^{|B_L|}$, probably a really really big number.

Nonetheless, reasonable-size machines exist for many hyperbolic groups. See, for instance, figure 11.10, Pentagon word acceptor, for a diagram of the word acceptor for the group generated by reflections in the sides of a right-angled pentagon in the hyperbolic plane.

It is not hard, in general, to fix up an automatic structure so that it is prefix-closed, or to fix up the structure so that it represents each element uniquely. What is hard is to find a general procedure which will do both simultaneously, although I do not know any example of a group which admits an automatic structure but does not admit one which is prefix-closed and unique.

There is a strong connection between automatic structures on the fundamental groups of compact hyperbolic manifolds and orbifolds and self-similar tilings of the plane. In fact, the geometry of similarities of the Euclidean plane is closely linked to hyperbolic geometry: if G is the group of similarities, then G/K is homeomorphic to $\mathbf{R}^3$, where K is a maximal compact subgroup, namely the group $SO(2)$ of rotations about a point. G/K can be given a metric which is invariant by the action of G. The best such metric makes it isometric to $\mathbf{H}^3$. Thus G is a subgroup of the group of isometries of $\mathbf{H}^3$: the subgroup which fixes the point at infinity in the Poincaré upper half space model. Geometrically, if one paints a pattern on the bounding plane, and looks 'down' at this plane while moving around in $\mathbf{H}^3$, one sees the pattern shrinking as one goes higher, expanding as one goes lower — the view transforms by similarities.

Related to this, if one has a scheme for self-similar tilings of the plane, one can make three-dimensional hyperbolic blocks which encode the rules. Choose a horosphere h_1 (in the upper half-space model, a good choice is a horizontal plane at height 1.) Make a copy of each tile type on this plane. Let h_2 be the horosphere which has hyperbolic distance $\log(|\lambda|)$ outward from h_1, where λ is the expansion constant. In the upper half-space model, this would be the horizontal plane at height $1/\lambda$. For each tile, form a solid block by sweeping the tile down, each point on the tile following a geodesic perpendicular to the two horospheres, until it meets the second horosphere. The outer face (on h_2) is expanded by a factor of $|\lambda|$. On the lower horosphere, paint the pattern of the subdivision of the tile.

A self-similar tiling or 'almost self-similar tiling' of the plane in this way generates a tiling of hyperbolic 3-space, which incorporates at once the tiling at all scales. The tiling of $\mathbf{H}^3$ has a natural spanning tree or forest, which connects each parent tile to its children through their mutual horospherical faces. This tree is recognized by a finite state automaton, just as the tree of Proposition 11.11, hyperbolic automatic tree. In fact, in some cases, the the two constructions give combinatorially identical trees beyond a certain point.

Similarly, the automatic structures on hyperbolic groups give tilings of the sphere at infinity in hyperbolic space: the sphere can be divided up into a finite number of pieces according to which depth k branch of the tree feed it. These tilings are not self-similar — indeed, the sphere has no similarities — but they are eventually 'self-Moebius'.

There are some further results on existence:

PROPOSITION 11.12. FINITE INDEX AUTOMATIC. *A group which contains an automatic group of finite index is itself automatic. A subgroup of finite index in an automatic group is automatic.*

PROPOSITION 11.13. PRODUCT AUTOMATIC. *A product or free product of a finite number of automatic groups is automatic.*

THEOREM 11.14. CENTRAL EXTENSIONS AUTOMATIC. *If H is a hyperbolic group, A is an abelian group, and*

$$A \to G \to H$$

is a central extension, then G is hyperbolic.

Here a hyperbolic group can be taken to be the fundamental group of a compact, negatively curved, orbifold with convex boundary, or (possibly more generally), a hyperbolic group in the sense of Gromov.

COROLLARY 11.15. AUTOMATIC NOT NON-POSITIVE. *There are closed 3-manifolds which do not admit metrics of non-positive negative curvature whose fundamental groups are automatic.*

Any fiber bundle (or Seifert fiber space) over a closed surface has an automatic fundamental group; most of these do not admit metrics of non-positive curvature. The construction is related to the fact that the metrics on their universal covers are quasi-equivalent to metrics of non-positive curvature.

The condition that H be not only automatic but hyperbolic is essential on account of the following examples:

THEOREM 11.16. NILPOTENT GROUPS NOT AUTOMATIC (HOLT). *A nilpotent group is automatic if and only if it contains an abelian subgroup with finite index.*

THEOREM 11.17. BRAID GROUPS AUTOMATIC. *The braid groups have automatic structures*

THEOREM 11.18. SL(N,Z) NOT AUTOMATIC. *The groups $SL(n, \mathbf{Z})$ are not automatic for $n \geq 3$. In fact, the graphs of these groups do not admit combings.*

CONJECTURE 11.19. NONPOSITIVE NONAUTOMATIC. *A cocompact group of isometries of $\mathbf{H}^2 \times \mathbf{H}^2$ which is not an almost product of surface groups is not automatic.*

This conjecture, if verified, would show that the condition for a group to be automatic depends not just on the quasi-geometric of the group, but on combinatorial properties as well — since the graph of any such group is quasi-equivalent to the graph of the product of two surface groups, which is automatic.

REFERENCES

[Conway Lagerias] J.H. Conway and J.C. Lagarias, *Tiling with Polyominoes and Combinatorial Group Theory.*

[CEHPT] J. Cannon, D. Epstein, D. Holt, M. Paterson, and W. Thurston, *Word processing and group theory*, Preprint.

[Lind] D. Lind, Bull. AMS.

[Knuth] Donald E. Knuth, "The Art of Computer Programming."

[Milnor Thurston] J.W. Milnor and W.P. Thurston, *On iterated maps of the interval*, in "Dynamical Systems," Springer-Verlag, 1988.

[Thurston0] William .P. Thurston, *Conway's tiling groups*, American Mathematical Monthly an 1990.

[Thurston1] William P. Thurston, "The Geometry and Topology of 3-Manifolds."