THE MINKOWSKI QUESTION MARK AND THE MODULAR GROUP $PSL(2,\mathbb{Z})$ (EXPOSITORY)

LINAS VEPSTAS

ABSTRACT. Fractals and continued fractions seem to be deeply related in many ways. Farey fractions appear naturally in both. Much, maybe even most of this relationship can be explained by the fact that both are represented with the infinite binary tree. In turn, this tree describes the structure of the Cantor set.

This text is meant to be an easy-to-read and informal presentation as to how all of these ideas tie together, to reveal how they are all just different representations of the same thing. Reading this text does not require any formal education in mathematics: it should be accessible to anyone who is interested. That said, a raft of fancy keywords provide hooks for finding more information via search engines. Here's the raft.

The infinite binary tree can be viewed as a certain subset of the modular group $PSL(2,\mathbb{Z})$. This is the dyadic groupoid or dyadic monoid. It provides a natural setting for the symmetry and self-similarity of many fractals, including those associated with period-doubling maps, with phase-locking maps, and with various dynamical systems in general.

The dyadic monoid appears to be isomorphic to a large collection of other things: the rational numbers; the dyadic rationals (have a power of two in the denominator), the infinite binary tree, the Cantor space, viewed as a set of infinitely long strings of 1's and 0's, the Farey and Stern-Brocot trees, continued fractions, the set of quadratic irrationals, Baire space (viewed as a set of infinitely long strings of positive integers). The dyadic monoid has a surjection onto the real numbers. When also passing through the continued fractions, this gives the Minkowski Question Mark function. All of these inter-relationships are presented in this text.

A particular amount of effort is expended on demonstrating the fractal self-similarlity, the symmetry properties of all of these objects. This is examined from many different angles, so as to make clear it's "always the same thing".

XXX Caution: this paper is in a perpetual state of being unfinished. This version corrects a number of serious errors found in previous drafts. Yet, it will surely still be misleading and confusing in many ways. The second half, in particular must surely contain errors and mis-statements! Caveat emptor! XXX

1. INTRODUCTION

The Minkowski Question Mark function, shown in figure 1.1, has many strange and unusual properties. As is readily apparent, it is continuous everywhere, and monotonically increasing. A naive attempt to compute its derivative seems to show that it has a derivative that is zero "everywhere", or at least, zero on the rational numbers. A visual examination shows that it is clearly self-similar, yet the self-similarity cannot be a simple re-scaling, as some stretching and shrinking is needed to make the self-similarity work. The goal of this paper is to describe, in simple terms, its analytic and topological properties, and, most

Date: 12 October 2004 (updated 22 August 2014).
December 2023: Bug fixes; significant rewrites of assorted sections.

FIGURE 1.1. The Minkowski Question Mark Function

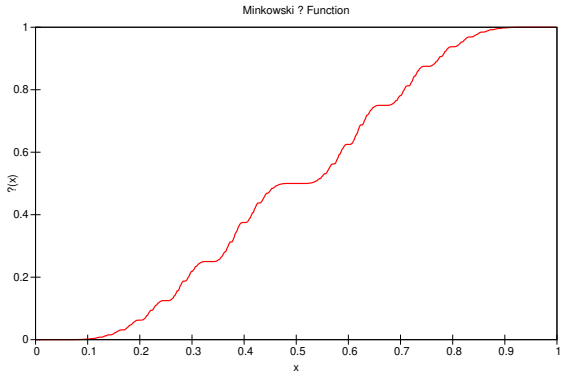
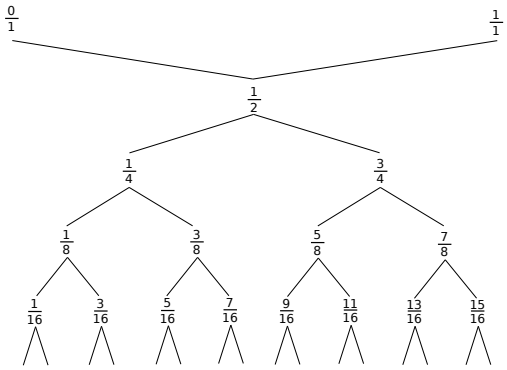


FIGURE 1.2. The Dyadic Tree



importantly, its self-similarity properties. These will be seen to be the self-similarities of the infinite binary tree.

1.1. **The Binary Tree.** The infinite binary tree is depicted in figure 1.2. Its clearly "binary" in that branching downwards, there are two offshoots from each node. The nodes themselves are labeled with binary numbers, or, more precisely, with "dyadic fractions". A dyadic fraction is simply a fraction whose denominator is a power of two. This text will distinguish between binary and dyadic, with "binary" referring to the shape of the tree, and "dyadic" referring to powers of two. For example, the nodes of the binary tree need not be labeled with dyadic fractions. When they are, it may be called the "dyadic tree".

The binary tree is readily navigated when descending down it, by making a sequence of left and right moves. Labeling these moves by L and R, every node can be uniquely labeled by the path taken to get to it from the root of the tree. A general node of the tree thus corresponds to some string of letters $L^m R^n L^p \cdots$ for some non-negative integer m and positive integers n, p, ... The superscript, as usual, simply means that a given letter is

repeated some number of times, so that

$$L^m R^n L^p \cdots = \underbrace{LLL\cdots L}_{m} \quad \underbrace{RR\cdots R}_{n} \quad \underbrace{LL\cdots L}_{p} \quad \cdots$$

For the dyadic tree, the path label can be directly converted to the dyadic label: the string of L's and R's can be taken as a string of 0's and 1's, a binary expansion; one adds an extra 1 at the end to get the dyadic value. Thus, starting at the root of the tree, taken to be $1/2$, a series of left and right moves takes one to the following nodes:

$$\begin{array}{ccccc} L & = & 0.01 & = & 1/4 \\ R & = & 0.11 & = & 3/4 \\ RL & = & 0.101 & = & 5/8 \\ L^2 & = & 0.001 & = & 1/8 \\ L^2RL & = & 0.00101 & = & 5/32 \end{array}$$

Selecting a node in the tree is the same as selecting a subtree, in that the node is the root of the entire subtree underneath it. Any subtree is clearly isomorphic to the whole tree, and it is from this property that self-similarity follows for any system that can be mapped onto the binary tree. If the nodes of the tree are labeled in strictly ascending order, as they are in the dyadic tree, then selecting a node in the tree is the same as specifying an interval: the interval runs from the lim *inf* to the lim *sup* of the subtree. The converse is not true: a general interval will not correspond to a single subtree. One may choose the lower limit of an interval arbitrarily, but the upper limit of the interval will be constrained by the possible subtrees with the given lower limit. These ideas, of the equivalence of intervals and trees and nodes, will be made more precise in later sections.

1.2. **The Farey Tree and the Stern-Brocot Tree.** The Farey tree[1], at whose nodes sit the Farey fractions[2], is depicted in figure 1.3. It is a binary tree labeled with rational numbers in the most peculiar fashion, and the tree has many unusual and interesting number-theoretic properties. It is constructed by means of mediants. The mediant of two fractions p/q and r/s is defined as $(p+r)/(q+s)$: one adds numerator and denominator, as if making a school-child mistake. One begins the construction by labeling the end-points of the unit interval as $0/1$ to $1/1$, and arranging them into a row, the zeroth row: $\left\{\frac{0}{1}, \frac{1}{1}\right\}$. The first mediant is $(0+1)/(1+1) = 1/2$, which is placed in the middle to create the first row: $\left\{\frac{0}{1}, \frac{1}{2}, \frac{1}{1}\right\}$. At the next iteration, one may construct two more mediants,

$$\begin{array}{ccc} \dfrac{(0+1)}{(1+2)} & = & \dfrac{1}{3} \\ \dfrac{(1+1)}{(2+1)} & = & \dfrac{2}{3} \end{array}$$

which are placed in between their progenitors, so: $\left\{\frac{0}{1}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{1}{1}\right\}$. The construction proceeds onward in this manner, with the mediant being taken of neighboring fractions. The tree itself is obtained very simply by placing the mediants onto the the respective positions in the tree.

Neighboring fractions in each row have the curious property of being "unimodular": by this, it is meant that if p/q and r/s are neighboring fractions, then $rq - ps = 1$. This is easily proved by induction: clearly, the relation holds for $0/1$ and $1/1$. One then shows that if the relationship holds for the pair $\left(\frac{p}{q}, \frac{r}{s}\right)$, then it also holds for the two pairs $\left(\frac{p}{q}, \frac{p+r}{q+s}\right)$

FIGURE 1.3. The Farey Tree

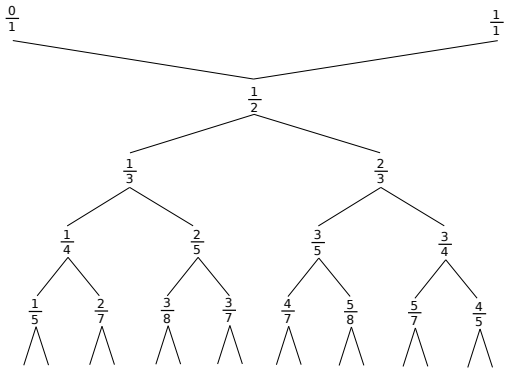
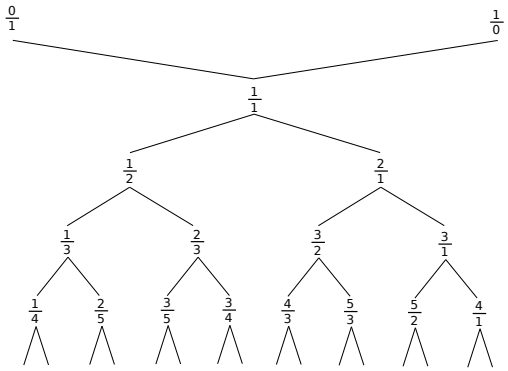


FIGURE 1.4. The Stern-Brocot Tree



and $\left(\frac{p+r}{q+s}, \frac{r}{s}\right)$. The term "unimodular", while seemingly opaque here, is no accident. The pair of fractions, written as a 2x2 matrix

$$A = \begin{bmatrix} r & p \\ s & q \end{bmatrix}$$

has a unit determinant, and so $A \in SL(2,\mathbb{Z})$, where $SL(2,\mathbb{Z})$ is the group of 2x2 matrices with unit determinant. Although its appearance here seems arbitrary, this group of matrices will recur more deeply throughout the theory of binary trees. The unimodular property then allows one to prove that

$$\frac{p}{q} < \frac{p+r}{q+s} < \frac{r}{s}$$

and so mediants always lie strictly in between their progenitors. This implies that every row of Farey fractions are in strict ascending order. From this it follows that any given fraction can appear only once in a given row, and that the progenitors of a given fraction are unique. In fact, it may be shown that every fraction appears somewhere in the tree; several simple proofs are provided in [1].

Closely related to the Farey tree is the Stern-Brocot tree, depicted in figure 1.4. The construction proceeds in a similar manner to that of the Farey tree. The primary difference is that the Stern-Brocot tree spans the entire non-negative real number line $0/1 \leq p/q \leq 1/0 = \infty$, instead of just the closed unit interval. As is readily apparent from the figure, the left half of the tree is identical to the Farey tree. Equally apparent, the right half has the same form as the left half, but with the fractions p/q turned upside-down, to form q/p. Matching up the nodes of the Farey tree to the nodes of the Stern-Brocot tree induces a function from the the rationals on the unit interval to all positive rationals. This function is given by

$$f(x) = \frac{x}{1-x} \tag{1.1}$$

It is not hard to imagine that this function holds for real numbers as well: that one can take not only an unbounded number of steps down the tree, but that one can go infinitely far, arriving at the real numbers. The reals can be imagined as those places that correspond to the gaps between the branches in the tree. That is, one can imagine that chasing a single branch all the way down will still arrive at something that "feels like" a rational number. The reals, somehow, live "in between". The function $x/(1-x)$ does "make sense" for the reals as well as the rationals. The limit seems to be well-defined, the function seems continuous.

This will be almost the last time that the word "continuous" is mentioned in this text. For two reasons. First is that this is not really the topic that this text wants to talk about. Second is that it really can't: a proper, formal defintion of continuity requires the defintion of open sets, and, more generally a topology. This requires a lot of work, and if one is not careful to bound the discussion, it can reach into the depths of general topology and descriptive set theory. Entire books have been written on these topics, and more than a few. There is no topology in this text. Most of the discussion will be about single points, individual instances of things: perhaps integers, perhaps strings of integers, perhaps rationals.

Another characteristic of this text is that it will carefully avoid using any kind of sophisticated mathematics at all. There will be a number of quite fancy words brandished here or there. These can be used in a search engine, to find more information. There will be no appeals to any fancy or deep theorems from mathematics. This text is meant to be readable by enthusiasts and amateurs. It is not aimed at mathematicians.

1.3. **The Question Mark Function.** By identifying the dyadic tree and the Farey tree, one obtains the Minkowski question mark function as the map between the two trees. The question mark function is denoted by $?(x)$, and is the map of labels from the Farey tree to the dyadic tree, so that, for example, $?(1/3) = 1/4$. The recursive construction of the two trees also allows a recursive construction of the question mark function[3, Contorted fractions, chapter 8]. At the endpoints, one has $?(0) = 0$ and $?(1) = 1$. Then, given a pair of neighboring progenitors p/q and r/s, one equates the Farey mediant to the arithmetic average:

$$?\left(\frac{p+r}{q+s}\right) = \frac{1}{2}\left[?\left(\frac{p}{q}\right) + ?\left(\frac{r}{s}\right)\right] \tag{1.2}$$

That this gives a node on the dyadic tree follows easily, as the dyadic tree may be constructed in the same way that the Farey tree was, with arithmetic average taking the place of the mediant.

The recursive construction provides a map from the rationals to the dyadic rationals. Given a square-free number p, define

$$\mathbb{Q}_p = \left\{ \frac{m}{p^n} \middle| n \in \mathbb{N}, 0 \le m < p^n \right\} \tag{1.3}$$

to be the p-adic rationals on the unit interval, that is, the rational numbers with a power of p in the denominator. Here, $\mathbb{N}$ denotes the set of natural numbers. The term "square-free" simply means that all of the prime factors of p occur only once; p contains no squares. With this notation, the question mark function is then a map

$$?: \mathbb{Q} \to \mathbb{Q}_2$$

where, by abuse of notation, $\mathbb{Q}$ is understood to be the rational numbers on the unit interval. The restriction to the unit interval avoids the question of how to extend the question mark function to larger positive or negative arguments; several different, inequivalent extensions are possible. In all of what follows, whenever the symbols $\mathbb{Q}$ or $\mathbb{Q}_2$ or even $\mathbb{R}$ are used, these should be understood to be limited to the unit interval; this avoids the need for a more cumbersome notation.

The rationals $\mathbb{Q}$ and the dyadics $\mathbb{Q}_2$ are both dense in the reals $\mathbb{R}$. A well-known theorem of general topology states that any continuous map of dense subsets of $\mathbb{R}$ can be uniquely extended to a continuous map for all of $\mathbb{R}$. It is straight-forward to show that the question mark function is continuous in the usual topology on the unit interval: given any open interval (a,b), the preimage $?^{-1}(a,b)$ is also an open set. The proof of continuity follows from the fact that both the dyadic and the Farey trees were strictly ordered (if $p/q < r/s$ then p/q will appear to the left of r/s in the tree), and that all possible dyadics appear in the dyadic tree, and all possible rationals appear in the Farey tree. Thus, one may conclude that the question mark function is well-defined on the real numbers, and furthermore, that it is monotonically increasing.

1.4. **Representations of Real Numbers.** An alternate definition of the question mark function may be given that is more suitable for many practical computations, than the recursive definition of equation 1.2. The alternate definition is given in terms of continued fractions, and requires a brief diversion into the representation of real numbers.

Let $x = [a_1, a_2, \cdots]$ be the continued fraction representation[4] for a real number x. In this representation, one has

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cdots}}}$$

for positive integers $a_1, a_2, \cdots$. The word "representation" here is not lightly chosen, but is meant to evoke a deeper idea: that the real numbers exist as an abstract set, whereas the continued fraction is one way of writing down the value of a real number in a manipulable way. Other representations are possible: of course, everyone is taught the decimal expansion (the base-10 or 10-adic representation). Formally, one writes

$$[\,]: \mathbb{R} \to \mathbb{N}^\omega$$

where $[\,]$ is understood to be the operation of creating the continued fraction expansion of a real number, and $\mathbb{N}^\omega = \mathbb{N} \times \mathbb{N} \times \cdots$ is understood to be the Cartesian product of a countable infinity of copies of the natural numbers $\mathbb{N}$. The space $\mathbb{N}^\omega$ is sometimes called Baire space; however, the standard definition of Baire space comes with a natural topology on it, and we have not yet broached the subject of topologies.

By contrast, the p-adic or base-p representation of a real number is given by

$$(\,)_p : \mathbb{R} \to \mathbb{Z}_p^\omega$$

Here, $\mathbb{Z}_p = \{n | n \in \mathbb{Z}, 0 \leq n < p\}$ is the set of integers from 0 to $p-1$, and so $\mathbb{Z}_p^\omega = \mathbb{Z}_p \times \mathbb{Z}_p \times \cdots$ is the product of a countable infinity of copies of $\mathbb{Z}_p$. Here, the symbol $(x)_p$ just means, very simply, "take the base-p expansion of the real number x", so that, for example, $(\pi)_{10} = 3.141592653\cdots$.

A remarkable and sometimes-forgotten property of the p-adic representation is that it is not isomorphic to the real numbers. A common school demonstration is that $0.9999\cdots = 1.000\cdots$: there are two different base-10 expansions that are equal to the same real number. In fact, this ambiguity exists for any p-adic fraction when written out in base-p. This happens for *every* p-adic representation. There is a very simple way of visualizing this problem in terms of trees. Consider, for example, $p = 2$. A 2-adic fraction corresponds to a node in the dyadic tree:the dyadic tree provides a representation for the 2-adic fractions. Consider now the real number obtained by starting at a given node, taking the left branch, and then a succession of right branches, as so: *LRRRR*$\cdots$. after an infinite number of steps, one arrives at a "leaf" of the tree. The numerical value of the leaf, expressed as a real number, is identical to the starting node. Similarly, one may take the right branch, followed by a succession of left branches, like so: *RLLLL*$\cdots$ One concludes that the dyadic tree, as a representation of the real numbers, is triply degenerate at the 2-adic fractions, in that $\frac{1}{2} = 0.1000\cdots = 0.01111\cdots$. In fact, the leaves of the dyadic tree form the Cantor set, a property that will be explored in a later section.

1.5. **Topology.** The space $\mathbb{N}^\omega$ has a natural topology, called the product topology, and when given that topology, is called Baire space. Likewise, the Cantor set $\mathbb{Z}_2^\omega$ or, more generally, $\mathbb{Z}_p^\omega$ also has the product topology as its "natural" topology. Neither of these topologies are isomorphic to each other, or the natural topology on the reals; it is the interplay between these that leads to many of the interesting properties discussed here.

1.6. **A Direct Form for the Question Mark Function.** For many purposes, including numerical exploration, a non-recursive definition of the question mark function is convenient to have. That is, given any rational or real x, one wants to be able to directly evaluate $?(x)$. This may be easily done by the use of continued fractions.

Given a continued-fraction expansion $x = [a_1, a_2, \cdots]$ of a number real number x, one then has an expansion given by Conway[3] (sometimes called the "Denjoy expansion"):

$$?(x) = 2\sum_{k=1}^{N} (-1)^{k+1}\ 2^{-(a_1+a_2+\cdots+a_k)} \tag{1.4}$$

where N is the length of the continued fraction; $N = \infty$ for irrational numbers. This sum can be visualized as a count of an alternating sequence of 0's and 1's in the binary expansion of $?(x)$:

$$?(x) = 0.\underbrace{000..0}_{a_1-1}\ \underbrace{11..1}_{a_2}\ \underbrace{00..0}_{a_3}\ \underbrace{11..1}_{a_4}\ \underbrace{00..0}_{a_5}1..$$

When N is finite, then the expansion is completed by an infinite repetition of the opposite of the last digit, so that, for N even, the binary expansion trails off with all 0's, while for N odd, it ends with all 1's. This definition enables a simple algorithm for the direct evaluation of both the question mark function, and its inverse: in one direction, one simply computes the continued fraction expansion of a real number, and converts it to a sum. In the opposite

direction, one simply counts digits in the dyadic expansion, using these to reconstruct the continued fraction.

Theorem. *The recursive definition of the Minkowski question mark function, given in equation 1.2, is equivalent to the direct definition of equation 1.4.*

Proof. To show equivalence, one needs to show that, after a sequence of left and right moves on the binary tree, the direct definition, in terms of continued fractions, gives a node on Farey tree labeled with the correct value of the Farey fraction. This equivalence may be made by induction.

The proof requires the use of partial convergents of continued fractions. Given a continued fraction $[a_1, a_2, \cdots]$, one defines the k'th convergent as

$$\frac{p_k}{q_k} = [a_1, a_2, \cdots, a_k]$$

that is, as the result of terminating the continued fraction at the k'th term. These are called convergents or approximants because these converge to the final value of the continued fraction:

$$[a_1, a_2, \cdots] = \lim_{k\to\infty} \frac{p_k}{q_k}$$

The numerator and denominator of the convergent obey well-known recursion relations[4]:

$$\begin{aligned} p_k &= a_k p_{k-1} + p_{k-2} \\ q_k &= a_k q_{k-1} + q_{k-2} \end{aligned}$$

A walk to the left or to the right on the dyadic tree is equivalent to appending a 0 or a 1 to the dyadic expansion of a number. Using the direct definition, the operation of left and right moves on continued fractions may be discerned. A left move is given by

$$L([a_1, a_2, \cdots, a_N]) = \begin{cases} [a_1, a_2, \cdots, a_N + 1] & \text{for } N \text{ odd} \\ [a_1, a_2, \cdots, a_N - 1, 2] & \text{for } N \text{ even} \end{cases}$$

whereas the right move is defined similarly, with the role of even and odd reversed. That is, a left move either increases the trailing entry in the continued fraction by one, or it inserts a comma, to start a new entry for the continued fraction. The seemingly strange appearance of the "-1,2" in the last is simply a byproduct of the identity $[a_1, a_2, \cdots, a_N] = [a_1, a_2, \cdots, a_N - 1, 1]$. That is, the left move adds a one to the last entry, but only after normalizing the continued fraction so that it has an odd length. The right move does the same, after first normalizing to an even length. Thus, to understand the general case, its sufficient to contemplate the value of $[a_1, a_2, \cdots, a_N + 1]$.

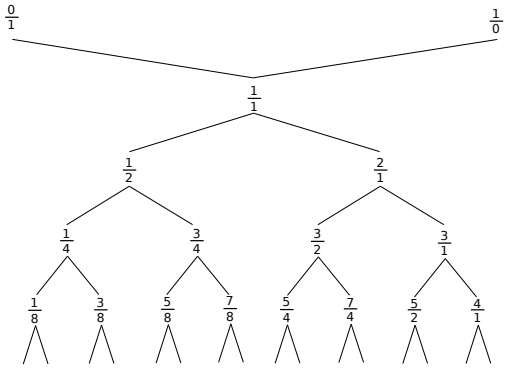
Let

$$\frac{r_N}{s_N} = [a_1, a_2, \cdots, a_N + 1]$$

be the convergent of the (left or right) subnode. One easily finds that $r_N = p_N + p_{N-1}$ and $s_N = q_N + q_{N-1}$. That is, the continued fraction $[a_1, a_2, \cdots, a_N + 1]$ is the mediant of its two last convergents, p_N/q_N and p_{N-1}/q_{N-1}. This provides the inductive step. The convergent p_N/q_N corresponds to the immediate predecessor of r_N/s_N in the Farey tree, while the convergent p_{N-1}/q_{N-1} represents the last "change of direction" in the left-right traversal. □

Just as the mediants of the Farey fractions were unimodular, so are the convergents of a continued fraction: $q_k p_{k-1} - p_k q_{k-1} = (-1)^k$, so that, as before, these may be arranged into a 2x2 matrix with unit determinant.

FIGURE 1.5. The Birthday Tree



1.7. **Extending to the Real Number Line.** The question mark function, as constructed above, is defined only on the unit interval. There are several different ways of extending the function to entire positive real axis. Ideally, this extension should have some sort of "natural" interpretation of a map from one tree to another. The Stern-Brocot tree, shown in figure 1.4, ranges over the entire positive real axis $(0,\infty)$. How should the dyadic tree 1.2 be extended to the same range? One such extension is known as the "Birthday Tree", and is shown in figure 1.5. Note that the birthday tree has the dyadic tree as the left subtree, and that the dyadic tree is repeated, shifted by one, under each positive integer. The positive integers appear along the right side of the tree. The birthday tree occurs as a very natural construction of the real numbers, given by J. H Conway, as an extension of the construction of the Peano arithmetic[3].

Aligning the birthday tree with the Stern-Brocot tree results in the extension

$$?(x) = \lfloor x \rfloor + ?(x - \lfloor x \rfloor)$$

where $\lfloor x \rfloor$ is the largest integer less than or equal to x. The left-handed side is defined for the entire positive number line, making use only of the definition of $?(x)$ on the unit interval. For comparison, the map between the dyadic tree and the birthday tree is given by

$$b(x) = \begin{cases} 2x & \text{for } 0 \le x \le \frac{1}{2} \\ 4x-1 & \text{for } \frac{1}{2} \le x \le \frac{3}{4} \\ \dots & \dots \\ 2^{n+1}x + n + 2 - 2^{n+1} & \text{for } \dfrac{2^n-1}{2^n} \le x \le \dfrac{2^{n+1}-1}{2^{n+1}} \end{cases}$$

The map from the Farey tree to the Stern-Brocot tree was given earlier, and is much simpler:

$$f(x) = \frac{x}{1-x}$$

One can deduce that $f = ?^{-1} \circ b \circ ?$ by lining up each of the trees. Formal tools that avoid the need for visual lining-up are developed in later sections.

1.8. **Pellian Equations.** The question mark function has the curious property that it maps quadratic irrationals to rational numbers, and thus arises indirectly in the theory of Pellian

equations. The connection can be demonstrated as follows. It has already been noted that the question mark maps rational numbers to dyadics. Dyadic rationals have, by definition, a finite length binary expansion; just as rational numbers have a finite number of terms in their continued fraction expansion: the question mark maps finite sequences to finite sequences. But what about periodic sequences? The fraction 1/9 has a repeating 2-adic expansion: $1/9 = 0.000111011101110\cdots$ Thus, the corresponding continued fraction must be periodic as well: $?([4,3,1,3,1,3,1,\cdots]) = 1/9$. What is the value of the continued fraction? Writing $[4,3,1,3,1,3,1,\cdots] = [4,x]$, the value of x must satisfy $x = [3,1,x]$ or $x = -1/2 + \sqrt{7/12}$, so that $[4,3,1,3,1,3,1,\cdots] = 1/(4+x) = \left(21 - \sqrt{21}\right)/77$. In other words, one has $?\left(\left(21 - \sqrt{21}\right)/77\right) = 1/9$.

The solution took the form $\left(j + k\sqrt{D}\right)/m$ for some integers j,k and positive integers D,m. A number of this form is called a quadratic irrational; every such number is a solution to a quadratic equation. Gauss had demonstrated that every continued fraction that eventually becomes periodic takes a value that is a quadratic irrational, and conversely, that every quadratic irrational has a continued fraction expansion that eventually becomes periodic. Gauss also demonstrated that the partial convergents of the continued fraction are solutions to an equation

$$p^2 - nq^2 = 1$$

for integer p,q and square-free integer n. This equation is the Pellian equation; clearly the rational p/q becomes a good approximation to $\sqrt{n}$. The appearance of the 1 on the right-hand side is again a manifestation of the unimodular relationship between partial convergents of the continued fraction: again, $SL(2,\mathbb{Z})$ enters the picture. Curiously, the classification of the solutions to the Pellian equation has a connection to the Riemann hypothesis.

A side effect of the question mark mapping is that it provides a simple proof that the set of quadratic irrationals is countable: each one is associated to a unique rational, and the rationals are countable.

XXX to-do: this section should be re-written to provide a better, stronger review of these (well-worn) topics.

2. Symmetries of the Question Mark

The Question Mark function is self-similar, and the generators of the similarity transformations can be written down explicitly. These generate a subset of a group (they do not generate a whole group, as will be made clear); a subset of the group $GL(2,\mathbb{Z})$. We'll give it a name: the *dyadic monoid*, although it can also be understood to be a semilattice; also, the dyadic monoid is a subset of the *dyadic groupoid*. The definitions of a monoid, a groupoid, and a lattice, will be reviewed later; the reason for naming the dyadic monoid now is to voice a hypothesis: the dyadic monoid gives the symmetry of all period-doubling fractal and chaotic phenomena.

Given a real number x having the continued fraction expansion $x = [a_1, a_2, \cdots]$, define the function $g(x)$ to be $g(x) = [a_1 + 1, a_2, \cdots]$. A straightforward manipulation shows that $g(x)$ has an explicit, concrete form: $g(x) = x/(x+1)$. Making use of the "direct" definition of the question mark 1.4, one has that g is a homomorphism of the question mark:

$$(?\circ g)(x) = ?(g(x)) = ?\left(\frac{x}{x+1}\right) = \frac{?(x)}{2} \tag{2.1}$$

Here, $\circ$ denotes function composition. The above identity follows, as adding one to $a_1$is the same as dividing by two in 1.4. By defining $h(x) = x/2$, the above identitiy can be written in the form of a commuting diagram: $?\circ g = h\circ ?$ which holds for all x in the unit interval.

The repeated application of the function g generates a symmetry of the question mark. Denoting repeated iterations by g^n, one readily obtains $g^n(x) = x/(nx+1)$. Equivalently, for a continued fraction,

$$g^n([a_1, a_2, \cdots]) = [a_1 + n, a_2, \cdots]$$

Iterating under the question mark gives

$$(?\circ g^n)(x) = ?\left(\frac{x}{nx+1}\right) = \frac{?(x)}{2^n} = (h^n\circ ?)(x)$$

The generator g maps intervals to intervals, specifically $g^n : [[0,1]] \to [[0, \frac{1}{n+1}]]$. Here the non-standard notation $x \in [[a,b]]$ is introduced to denote an interval $a \le x \le b$; the double bracket form is used only to avoid confusion with [] for continued fraction expansions. On the dyadic side, we have $(?\circ g^n) : [[0,1]] \to [[0, 1/2^n]]$.

The other symmetry of the question mark is more trivial: the question mark is symmetric under a left-right reflection:

$$?(1-x) = 1 - ?(x)$$

Denoting this by r for "reflection", a reflection operator may be defined as $r(x) = 1 - x$. As an operator, r commutes with ?, since $(?\circ r)(x) = (r\circ ?)(x)$. Clearly, r cannot be applied more than once, since r^2 is the identity. The reflection operator r, composed with g, generates more self-symmetries. Thus, for example,

$$\begin{aligned}(?rg^n)(x) = (r?g^n)(x) = (rh^n?)(x) &= 1 - \frac{?(x)}{2^n} \\ = ?\left(r\left(\frac{x}{nx+1}\right)\right) &= ?\left(\frac{(n-1)x+1}{nx+1}\right)\end{aligned}$$

shows that rg^n is the operation that, under ?, maps $[[0,1]] \to [[1, 1-1/2^n]]$. A general self-similarity transform may be written as

$$\gamma \equiv g^{a_1} r g^{a_2} r \ldots r g^{a_N} \tag{2.2}$$

Applying the two commutation relations $?g = h?$ and $r? = ?r$ repeatedly results in

$$\begin{aligned}(?\circ\gamma)(x) &= (?\circ g^{a_1} r g^{a_2} r \ldots r g^{a_N})(x) \\ &= (h^{a_1} r h^{a_2} r \ldots r h^{a_N} \circ ?)(x) \\ &= \frac{1}{2^{a_1}}\left(1 - \frac{1}{2^{a_2}}\left(1 - \frac{1}{2^{a_3}}\left(\cdots\left(1 - \frac{1}{2^{a_N}}?(x)\right)\right)\right)\right) \\ &= \frac{1}{2^{a_1}} - \frac{1}{2^{a_1+a_2}} + \frac{1}{2^{a_1+a_2+a_3}} - \cdots + (-)^{N+1}\frac{?(x)}{2^{a_1+a_2+a_3+\ldots+a_N}}\end{aligned} \tag{2.3}$$

To see the other leg of the commutative diagram of this homomorphism, observe that $g^n(x)$ may be itself written as a continued fraction:

$$g^n(x) = \frac{1}{n + \frac{1}{x}} = [n, x]$$

In the same vein, one has $(rg^n)(x) = [1, n-1, x]$ and, setting $n = 0$, with the slight abuse of allowing negative integers in continued fractions:

$$r(x) = \cfrac{1}{1 + \cfrac{1}{-1 + \cfrac{1}{x}}} = [1, -1, x]$$

Continuing this exercise, one finds that

$$\gamma(x) = (g^{a_1} r g^{a_2} r \cdots r g^{a_N})(x) = [a_1 + 1, a_2, a_3, \cdots, a_{N-1}, a_N - 1, x] \tag{2.4}$$

To obtain the explicit form of this as a fraction, one need only remark that both g and r take the form of a "fractional linear transform" or a "Möbius transform". That is, given a general 2x2 matrix, one defines the action

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} : x \to \frac{ax+b}{cx+d} \tag{2.5}$$

The utility of the fractional linear transform is that it commutes with matrix multiplication; one may multiply matrices on the left-hand side to get the correct expression on the right-hand side. In terms of Möbius transformations, one then has that

$$g = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad r = \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix} \tag{2.6}$$

and so $g^{a_1} r g^{a_2} r \ldots r g^{a_N}$ may be evaluated by matrix multiplication, followed by an application of the Möbius transformation 2.5 to obtain the fractional form.

2.1. **Useful Identities.** Some curious combinations of these operators are tabulated here. The action of r on a continued fraction is given by

$$r([a_1, a_2, ..]) = \begin{cases} [1, a_1 - 1, a_2, \cdots] & \text{for } a_1 \neq 1 \\ [a_2 + 1, a_3, \cdots] & \text{for } a_1 = 1 \end{cases}$$

Using the above, the operator to insert a digit $n \geq 1$ at the front of the continued fraction expansion is

$$(g^{n-1} r g)([a_1, a_2, \cdots]) = [n, a_1, a_2, \cdots]$$

which can be verified by using the Möbius transformation

$$(g^{n-1} r g)(x) = \left(\begin{bmatrix} 1 & 0 \\ n-1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \right) : (x) = \begin{pmatrix} 0 & 1 \\ 1 & n \end{pmatrix} : (x) = \frac{1}{n+x}$$

The repeated application of the insertion operator leads to

$$(g^{j-1} r g)(g^{k-1} r g)(g^{m-1} r g) \cdots (x) = (g^{j-1} r g^k r g^m r g \cdots)(x) = [j, k, m, \cdots, x]$$

which provides a simple proof of equation 2.4.

The append operator follows by combining

$$[a_1, a_2, \cdots, a_N, g^k(x)] = [a_1, a_2, \cdots, a_N + k, x]$$

and

$$[a_1, a_2, \cdots, a_N, r(x)] = [a_1, a_2, \cdots, a_N + 1, -1, x]$$

The appending operator is

$$\left(g^{-1} r g^{k+1} \right)(x) = \frac{kx+1}{x}$$

which, for $k \geq 1$, acts as

$$[a_1, a_2, \cdots, a_N, (g^{-1} r g^{k+1})(x)] = [a_1, a_2, \cdots, a_N, k, x]$$

Appending to the back, instead of the front, starting with $(g^{-1}rg)(x) = 1/x = [x]$ gives

$$(g^{-1}rg)(g^{-1}rg^{j+1})(g^{-1}rg^{k+1})(g^{-1}rg^{m+1})\cdots(x) = (g^{j-1}rg^krg^mrg\cdots)(x) = [j,k,m,\cdots,x]$$

which has the same form as before. One can append to the front or the back, and get the same results: the operation is associative.

2.2. **What is a Real Number?** In the above, the dyadic numbers and the continued fractions were casually taken as "being the same thing" as the real numbers. This section will attempt to be more precise.

Some formalities are in order. The set of all infinite strings of binary digits is called the Cantor space. It can be written as $2^\omega = \{0,1\}^\omega$ where $2 = \{0,1\}$ is just the set containing two items (two letters, which happen to be 0 and 1 at present). The symbol ω is the ordinal representing countable infinity. Thus, 2^ω is a product $\{0,1\} \times \{0,1\} \times \cdots$ of countably many copies of $2 = \{0,1\}$. Likewise, write $\mathbb{N}$ for the natural numbers, the set of positive integers. The countably infinite product $\mathbb{N}^\omega$ is called Baire space.

Both these spaces are well-known and carefully studied in mathematics; they have a large number of peculiar and interesting properties. None of this will be reviewed here: such a review would be longer than this text. For the present case, both will be treated as discrete spaces, just consisting of infinite strings of letters, and nothing more sophisticated. In particular, this section, and the rest of this text carefully avoids any discussions related to ideas from general topology. There will be no mention of open sets, nor of continuity. All objects should be thought of as points.

Both the Cantor space and the Baire space provide *representations* of the real numbers. What this means is that they both provide a way of writing down specific real numbers. The generic, abstract concept of a real number as some point on the real number line can be given a concrete and direct representation in terms of strings of integers. This enables a form of direct computation and manipulation that is not normally available in other theories that talk about the real numbers in their abstract setting.

It is well-known that representations of mathematical objects sometimes have additional properties that the "abstract objects" themselves do not have. There is an entire branch of mathematics devoted to this: this is model theory. This brings us to the primary point of confusion for this text: the dyadic monoid symmetries, already exhibited in the previous section, are these actual properties of the real numbers? Or are they purely a symptom of working with strings of symbols? Certainly, textbooks on pre-calculus, calculus or real analysis never-ever mention the dyadic monoid. Why is that?

Answering this last question is fairly easy, and can be traced back to Newton, calculus, and the development of analysis in the 17th, 18th and 19th centuries. There are three distinct steps. First is the recognition that there is a need for real numbers: there are things that simply aren't rationals, like $\sqrt{2}$, π and e. So, "real numbers exist". Next, the field of analysis developed from the need to formulate and solve differential equations. Either one works with "infinitesimals" (codified into "nonstandard analysis" in the 20th century) or one works with "delta-epsilon proofs", in order to avoid infinitesimals and some of their paradoxes by always working with finite quantities. To actually obtain a solution to a differential equation, it seems easiest to work with analytic expansions. Prototypically, this is the Taylor's series $f(x) = \sum_n a_n x^n$. Here, x is already a real number, and no longer a fraction. Differential equations can be solved by converting questions about derivatives acting on $f(x)$ into questions about the (infinite) series. This blooms into analysis, and eventually into the vast richness of modern mathematics.

The odd thing about analysis is that it (almost) never interacts with fractal structure. The real numbers of analysis are not fractals. They are things you can add, subtract, multiply, divide. They form a field. The fractal structure buried in Newton's method for approximating square roots was not appreciated until the 20th century. Is that fractal structure a property of the real numbers themselves, or is it "just" a property of Newton's algorithm? Perhaps it is just the algorithm, an infinitely recursive algorithm. Infinite recursion generically results in fractals.

The late 20th century saw the bloom of "chaos theory", when it was realized that differential equations have chaotic solutions, and that this chaos has something to do with fractals. Perhaps the omission of fractals from textbooks on real analysis is some grand historical mistake? It's not so simple. Lets review history. Things that can be added and multiplied turns into the ginormous branch of mathematics called "algebra", the field of real numbers occupies a distinguished spot. Analysis continues to be analysis, and the prototypical Taylor's expansion continues to smear out any fractal-like properties that might be possesses by real numbers. Cracks in the facade are explored with general topology, starting with separation axioms, moving through open sets, and making strong statements about metric spaces. Nothing in general topology makes use of, needs or exploits any sort of self-similarity properties. The exploration of symmetry is not a part of general topology. If reals have any sort of weird fractal property, general topology won't reveal it. How odd, given that general topology provides vast expanses of exotic, pathological examples.

One of the earliest explicit hints of the craziness of the real numbers arises with the Vitali set: one takes a quotient of the reals by the rationals, rearranges the points, and gets s set with no size. No appeal symmetry or self-similarity is required, other than to say "all cosets are the same" or perhaps "all rationals are the same", but saying that "these things are all the same" is just the feeblest venture into symmetry. The Banach–Tarski paradox takes a larger stride into symmetry, by employing the free group on two generators in it's statement. But again, the symmetry properties of this group remain unexplored and mostly unused.

There's a third direction in which history scoots off to. This is the development of set theory, the Peano axioms for arithmetic, ZF and ZFC, and the like. This cuts a bit closer to the heart of "what is a real number, really"? Apparently, some collections of axioms result in systems that do not have uncountable infinities in them: there might have "real numbers", but they are a bit different than the "standard" ones of the "standard model". Work on set theory does reveal collections of axioms that appear to be sufficient to provide everything needed for conventional analysis, and most of the rest of "standard" algebra and geometry. These axioms do allow for "real numbers" that behave, well, in completely "standard" ways. Yet, again, symmetry and self-similarity never show up for this party. There are at best some faint shadows: the idea of "well-foundedness" does require recursing to "the bottom of things". The idea of a power set does require considering all symmetric possibilities. Ultrafilters and ideals involve collections of "things that are like other things". But no appeal to any overt concepts of symmetry or self-similarity is needed to develop set theory.

Even descriptive "effective" set theory, which appeals to recursive algorithms as the foundation, never makes use of symmetry. Yes, recursive algorithms generically generate fractals. No, computatibility theory does not explore symmetry or self-similarity, for the most part.

This long discourse leaves us a bit empty handed in finding the answer to the question of "what is a real number, anyway?" The analysts have one answer that they like; the set

theorists have another. The topologists are happy with where they've arrived at: they have the "Polish spaces" and gaggles of theorems about and properties of Polish spaces.

The only ones who are worried are the chaos theorists, and they appear to be stumped. They interact with all of the above branches of mathematics, pulling theorems and formalisms as needed. However, to this day, there does not seem to be any grand. overarching, deep or intuitive explanation for why the Lorenz attractor is what it is. There's something fractal in there, and it involves real numbers in some way, but what, exactly, is it?

With that, we leave the question of what is a real number unanswered, and continue with an exploration of the symmetries of the Cantor space and of Baire space, both of which provide a *representation* of something that could be called "a real number", whatever that is. Whether these symmetries are actual properties of real numbers, or just properties of Cantor and Baire space will remain in the fog.

But still: a coda. If you dear reader, decide that these fractal self-similarities are merely properties of Cantor and Baire, and are not properties of "true real numbers", then ponder this: what explains the Lorenz attractor? Its a differential equation. It involves real numbers. How did the fractal properties sneak in there?

OK, OK, perhaps there's a hand-waving answer for that. Perhaps the Lorenz attractor acts on the space of all possible smooth functions in three variables, partitioning them into multiple pieces-parts, the so-called stable and unstable manifolds. Perhaps the correct statement is to say that the Cantor set shows up at the extreme boundary of these manifolds. They're kind-of knotted up, there; a Cantor knot at the boundary (What, exactly, is the knot theory of the Cantor set? Is it the Alexander horned sphere? The Cantor tree surface? Oh, never mind.). That's why things look fractal. But wait... what is the set of smooth functions in three variables? Doesn't this require real numbers, and real analysis, and topology, and more, to define the concept of "smoothness"? It all seems very circular, to me.

2.3. **Representations of Real Numbers.** Real numbers can be represented both with infinite strings of binary ones and zeros (the dyadic representation) and with infinite strings of integers (the continued fraction representation.) Neither representation is a bijection. For example, in the two strings $1000\cdots$ and $0111\cdots$ both represent the real number of 1/2. Looking at the the dyadic tree in figure 1.2, the first string corresponds to a move to the right branch, followed by an infinite choice of left branches. The second string corresponds to a move to the left branch, followed by an infinite choice of right branches. Perhaps this becomes more clear if one writes $RLLL\cdots$ and $LRRR\cdots$ instead of $1000\cdots$ and $0111\cdots$. But its the same thing: infinite strings of two symbols.

Such dyadic strings have a canonical mapping to the reals, given by

$$x = x_D = \sum_{k=1}^{\infty} b_k 2^{-k} \tag{2.7}$$

The meaning of the symbols is meant to be mostly obvious, but let's try to be precise, anyway. Each $b_k \in \{0,1\}$ is a bit; one is provided for each $k \in \mathbb{N}$. Here, x is a real number, but what is x_D? It is meant to be a *representation* of a real number, in the dyadic representation. A better notation can be provided by writing

$$\begin{aligned} x_D : 2^\omega &\to \mathbb{R} \\ b &\mapsto x = x_D(b) \\ x_D(b) &= \sum_{k=1}^{\infty} b_k 2^{-k} \end{aligned}$$

so that x_D is a function, a map, from 2^ω to $\mathbb{R}$. It maps strings b to binary expansions. The range of the map (the "codomain") is the unit interval of the reals, so $0 \le x \le 1$. But since, until now, we haven't defined any kind of topology on the reals, we haven't created any notion of open sets, it is not entirely appropriate to talk about an "interval". In particular, we don't even know what "less than" means for the reals, because we haven't demonstrated any kind of total order on them! Perhaps this all sounds very silly, since even elementary-school students have some intuitive feel for real numbers, and we're just being overly nit-picky. Whatever. Lets just fake it till we make it. Denote by $\mathbb{R}_D$ the range of the map x_D.

Note that every dyadic rational has two inequivalent representations in this map. A dyadic rational corresponds to a finite string, for which the b_k are defined only for $1 \le k \le N$. In this case, the dyadic rational is the number $m/2^N$ for some integer m (well, OK, to be explicit, its for the integer $m = \sum_{k=1}^{N} b_k 2^k$). Lets assume, without loss of generality, that $b_k = 1$; if it isn't, then just adjust N until it is. There are two infinite strings that map to this, corresponding to the left and right infinite branches on either side of the gap in the dyadic tree that lies underneath that dyadic rational:

$$b_R = (b_1, b_2, \cdots, b_{k-1}, 1, 0, 0, \cdots)$$
$$b_L = (b_1, b_2, \cdots, b_{k-1}, 0, 1, 1, \cdots)$$

These have $x_D(b_L) = x_D(b_R)$; they are the same real number. The map x_D is onto $\mathbb{R}_D$ but it is not one-to-one.

There's also a canonical map from infinite strings of natural numbers, to the reals. This is the continued-fraction mapping. As above, write

$$\begin{aligned} x_C : \mathbb{N}^\omega &\to \mathbb{R} \\ a &\mapsto x = x_C(a) \\ x_C(a) = x_C([a_1, a_2, \cdots]) &= \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cdots}}} \end{aligned}$$

As before, let $\mathbb{R}_C$ denote the range of this map. Note that the rational numbers are not a part of this map. A rational number corresponds to a finite-length string $[a_1, a_2, \cdots, a_N]$. So it would seem that $\mathbb{R}_C$ is contained in the unit interval, but is missing the rationals; it has holes. That is, x_C is injective (it is one-to-one) but it is not onto. We could try to patch this up by allowing some of the a_k to be zero. This results in $\infty = 1/0$ in the continued fraction, terminating an infinite string and giving a finite one back, which is identified as a rational. This works great, until one realizes that placing a zero in the sequence causes all the later ones to not matter. Yes, one gets a surjection, but there are infinitely many strings $[a_1, a_2, \cdots, a_N, 0, a_{N+2}, \cdots]$ that all map to the same rational. Ugh.

For completeness, it should be mentioned that there is a map between strings in $\mathbb{N}^\omega$ and the real unit interval that is a bijection, this is the beta map. However, it is far more complex to describe, and proving the bijection is not entirely easy. Details can be found in my other text on the beta map.[5]

Both of these representations are acted upon by the self-similarity transformation 2.2, but each in a different way. On the dyadic representation, the generator g has the representation

$$\begin{aligned} h &: \mathbb{R}_D \to \mathbb{R}_D \\ x &\mapsto x/2 \end{aligned}$$

when acting on the set of dyadic numbers, whereas as this same element has the representation

$$\begin{aligned} g \quad &: \mathbb{R}_C \quad \to \mathbb{R}_C \\ &\quad x \quad \mapsto x/(x+1) \end{aligned}$$

when acting on the set of continued fractions. The other generator, r, has the same representation on both sets: $r(x) = 1 - x$.

Define the set

$$M = \{\gamma \,|\, \gamma = g^{a_1} r g^{a_2} r \ldots r g^{a_N} \text{ for } a_k \in \mathbb{N} \text{ and } a_1, a_N \geq 0\} \tag{2.8}$$

Here, each a_k is a positive integer, with the exception of a_1 and a_N which may be zero. The set M is a *monoid*, and, more formally, is the *dyadic monoid*. Recall the definition of a monoid: it is a set with an associative operation (in this case, multiplication) that is closed under the operation, in the sense that if $\gamma_1 \in M$ and $\gamma_2 \in M$ then $\gamma_1 \gamma_2 \in M$. To complete the definition of a monoid, M must also have an identity element e, so that $e\gamma = \gamma e = \gamma$ for all $\gamma \in M$. There are no further axioms in the definition of a monoid.

Monoids are sometimes called *semigroups*, although the term semigroup is usually reserved for the idea of a monoid that does not have an identity element. The definition of a monoid is similar to that for a group, with an important difference: a monoid does not, in general, contain inverses. For a given $\gamma \in M$, there typically is no $\beta \in M$ such that $\gamma\beta = e$. For the present case, M cannot a group. Now, associative monoids, such as this, do have the universal property of being extendable to a full-fledged group, and one certainly could do this for M. However, such an extension can no longer be interpreted as a set of self-similarities acting on intervals. The extent to which some additional elements, that would act as inverses, can be added to M, is explored in a later section.

The role of the question mark function in connecting these two representations of the reals was just noted. This may be turned into a more formal theorem:

Theorem. *The Minkowski question mark provides a monoid homomorphism between $\mathbb{R}_C$ and $\mathbb{R}_D$. That is, there is a commuting diagram*

$$\begin{array}{ccc} \mathbb{R}_C & \xrightarrow{\gamma_C} & \mathbb{R}_C \\ ? \downarrow & \circlearrowright & \downarrow ? \\ \mathbb{R}_D & \xrightarrow[\gamma_D]{} & \mathbb{R}_D \end{array}$$

such that $? \circ \gamma_C = \gamma_D \circ ?$ holds for all $\gamma \in M$. More precisely, the homomorphism $(? \circ \gamma_C)(x_C) = (\gamma_D \circ ?)(x_D)$ holds for all $\gamma \in M$ and $x \in \mathbb{R}$, where $x_D \in \mathbb{R}_D$ and $x_C \in \mathbb{R}_C$ are both representations of the same real number x.

Proof. This will not actually be a prof, but yet more hand-waving. It summarizes and condenses some of the material in the earlier sections.

It was already noted, up above, that $? \circ g = h \circ ?$ and that $? \circ r = r \circ ?$ are homomorphisms on the rationals and dyadic rationals. This was made apparent by visual inspection and comparison of the Farey tree to the dyadic tree.

An abstract group element $\gamma \in M$ can be written in terms of the generators g and r as:

$$\gamma = g^{a_1} r g^{a_2} r \ldots r g^{a_N}$$

and clearly, $? \circ \gamma_C = \gamma_D \circ ?$ is arrived at by the finite number of applications of the two basic homomorphisms. Here, $\gamma_D = h^{a_1} r h^{a_2} r \ldots r h^{a_N}$, of course.

As before, the dyadic representation γ_D, acting on dyadic rationals x has the explicit form

$$\gamma_D(x) = \frac{1}{2^{a_1}} - \frac{1}{2^{a_1+a_2}} + \frac{1}{2^{a_1+a_2+a_3}} - ... + (-)^{N+1} \frac{x}{2^{a_1+a_2+a_3+...+a_N}} \tag{2.9}$$

The continued-fraction representation γ_C has the action

$$\gamma_C(x) = [a_1+1, a_2, a_3, ..., a_{N-1}, a_N - 1, x] \tag{2.10}$$

which holds for rational numbers x.

The task is to prove that $?(\gamma_C(x)) = \gamma_D(?(x))$.

The first part is silly-easy. If this equivalence holds for any given γ and γ' (at some fixed x), then it also holds for the product $\gamma\gamma'$ (at the same fixed x). By induction on γ, it holds for all $\gamma \in M$.

The second part is harder: prove that this holds for all x in the unit interval of the *reals*. This is the part where we throw our hands in the air, and give up. There are several distinct issues. One is that we haven't developed any notions of topology (of open sets), and thus have no notions of continuity, and thus cannot take limits. We can't show that some infinite sequence of rational numbers x_n converge to some limit $x_n \to x$ because, without some metric on the reals, we can't even show that the unit interval is bounded. Alas. Without boundedness and convergence, we've got no concept of continuity.

Even closure is hard to discuss: the rationals have the reals as a closure, but dyadics have another closure, the p-adics. These are "non-Archimedian" and have plenty of strange properties themselves.

Thus, we offer a proof by assertion: the question mark function is continuous. It is monotonically increasing, which means it makes sense to talk about it's inverse, and to treat the inverse as well-defined. In the grand scheme of things, the question mark is a continuous bijection on the real unit interval. Curiously, it is differentiable-nowhere; it is measurable, though. □

To summarize, we've made some progress, but not really. Somehow, the continued fractions $\mathbb{R}_C$ and dyadic numbers $\mathbb{R}_D$ are two distinct and inequivalent representations of the real numbers, both homomorphic to the abstract concept of the reals. These two representations are distinct, having different transformation properties under the action of the dyadic monoid. Clearly, $\mathbb{R}_D$ is homomorphic to the Cantor set, and $\mathbb{R}_C$ is homomorphic to Baire space. The Question Mark Function seems to have the general shape of something that could be an isomorphism between $\mathbb{R}_D$ and $\mathbb{R}_C$, both of which are isomorphic to the "true reals" $\mathbb{R}$ in the abstract. But we have none of the machinery needed to make this more precise. Indeed, there is a large literature on the Question Mark that clarifies and elucidates all sorts of issues involving continuity, smoothness, dimension, measurability, convergence, covering spaces and more. It's out there. For now, it is safe to just pretend, and continue onwards in a merry and naive way, treating the Question Mark as just some function on the real numbers (whatever those are).

2.4. **Blow-ups.** The self-similarity maps above were expressed in the form of mapping the whole interval to a sub-interval. Of course, the direction of this map can be turned around, so that sub-invervals are mapped to the whole interval. Things become interesting when one concatenates shrinks with blow-ups that are valid only on a sub-interval. This results in identities that are valid only on a sub-interval. However, the range of possibilities is much greater, and some of these identities take on new, unusual forms. Examples include

$$?(x)+\frac{1}{2}=?\left(\frac{1-x}{2-3x}\right) \quad \text{for } 0\le x\le\frac{1}{2}, \text{ having range } \frac{1}{2} \text{ to } 1$$

and

$$?(x)+\frac{1}{4}=\begin{cases} ?\left(\frac{1-2x}{3-7x}\right) & \text{for } 0\le x\le\frac{1}{3}, \text{ having range } \frac{1}{4} \text{ to } \frac{1}{2} \\ ?\left(\frac{4x-1}{5x-1}\right) & \text{for } \frac{1}{3}\le x\le\frac{1}{2}, \text{ having range } \frac{1}{2} \text{ to } \frac{3}{4} \\ ?\left(\frac{3-4x}{5-7x}\right) & \text{for } \frac{1}{2}\le x\le\frac{2}{3}, \text{ having range } \frac{3}{4} \text{ to } 1 \end{cases}$$

and so on.

XXX finish this section.

- Show how to construct these examples.
- All such examples derive from the form $\gamma=\alpha\beta^{-1}$ with $\alpha,\beta\in M$, and there are no others. This is proved as theorem 1 in section 5.8.
- The shrinking maps had continued fractions with positive entries – now consider continued fractions with negative entries. These maps have poles. But these poles never occur in the interval of validity.

3. Linear Groups and the Modular Group

Up to this point, we've been considering the infinite rooted binary tree, as shown in figures 1.2, 1.3, 1.4, 1.5, and the moves L and R on this tree, that walk one step down the tree, to the left and right, respectively. These moves can be concatenated: concatenation is a natural product structure on strings. Can we move upwards? Of course we can: it sort-of makes sense to talk about L^{-1} and R^{-1} as upwards moves on the tree. The inverse notation (the superscript -1) is suggestive and natural, as these are inverse moves: $L^{-1}L=LL^{-1}=RR^{-1}=R^{-1}R=e$ is the identity, as long as we are careful not to walk off the root of the tree. That is, we can almost understand L and R as generating a group, as long as we imagine ourselves starting somewhere in the middle of the tree, and don't walk so far so as to hit the root. Is it possible to do better? Is there a way of extending this structure so that its a real, full-fledged group? The answer is yes, and its easy and straightforward, once you see the way.

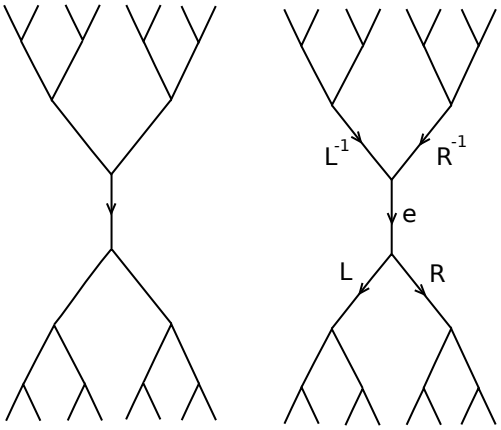
First, recall the definition of a group. It is a set of elements, with the following three properties:

- An associative product is defined on the set, such that the product ab of any two elements a and b always exists. This is called the 'closure' property. Associativity requires that $(ab)c=a(bc)=abc$.
- There is a special, unique element, the identity element called e, having the property that $ae=ea=a$ for any element a.
- Every element a has an inverse a^{-1}. The inverse is double-sided: $aa^{-1}=a^{-1}a=e$.

The issue preventing the moves L and R from forming a group is that, as one moves upwards in the tree, one eventually bumps into the root node. To make this clear, let's recall the definition of an action. An action is a pair (G,X) with X a set and G another set, possessing an associative product, and an identity element. The definition does not require that G have inverses, but that's OK if it does. If G is a group, then the action is called a *group action*. The action is defined with the following two statements:

- Elements $g\in G$ act on $x\in X$ such that $g\cdot x$ is another element in X. That is, $g\cdot x\in X$.
- The action is associative, so that $g\cdot(h\cdot x)=(gh)\cdot x$ for any two members $g,h\in X$.

FIGURE 3.1. Unrooted Tree



So, here, G is the set of moves on the tree, and X is the tree itself. The extension of moves on the tree to a full group can be achieved by simply eliminating the root of the tree! This can be done by gluing a second tree to the first, as shown in figure 3.1.

In the right-hand tree, the central edge is marked with an arrowhead. This singles it out as a special edge: the arrow is a device to keep track of one's position in the tree, as well as one's orientation, as one moves about. Without it, things can get confusing. The left hand tree shows the effect of the moves e, L, R, L^{-1} and R^{-1}. The identity e is, of course, no move at all. The four neighboring edges are marked with L, R, L^{-1} and R^{-1}; the arrowheads show how the central arrowhead travels, as the moves are applied.

Examining this figure, the moves L and R now clearly seem to be able to act without limitation, as the root no longer prevents an arbitrary length string of L^{-1} and R^{-1} from being applied. Thus, naively, it would seem that this simple trick allows the action of the free group $F_{\{L,R\}}$ in two letters $\{L,R\}$ to be defined on the unrooted infinite tree. However, something funny happens when one starts considering some non-trivial concatenations of moves. Consider, for example, the sequence of moves $LR^{-1}L$, illustrated in figure 3.2. This is essentially the same as the three-point turn performed in an automobile: one pulls forward L, turning, then back up R^{-1}, turning, and again pull forward L, straightening. The net result is that the car has been turned by 180 degrees!

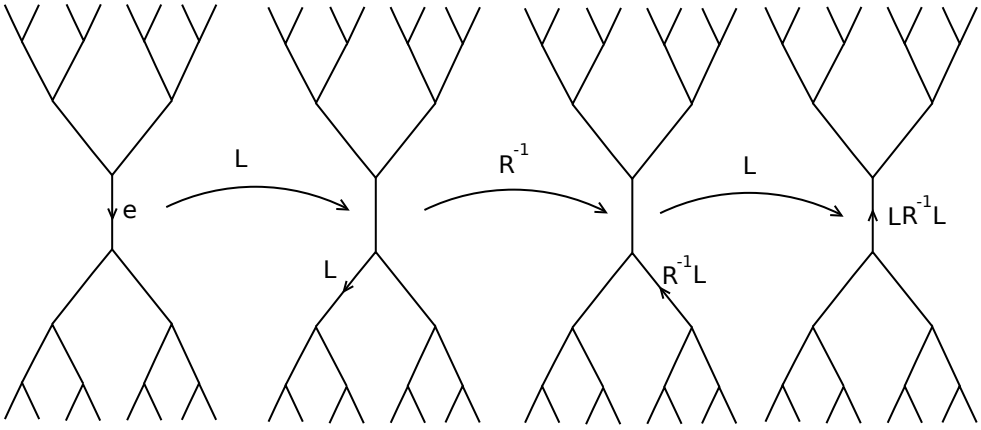
Of course, two such turns, back to back, rotate the car back into it's original position and direction. Thus, we conclude that, for the unrooted tree t, that there is a group action that is the identity $LR^{-1}L^2R^{-1}Lt = t$. We may interpret this formula in two ways: either we have the free group in two letters acting in an unfaithful, non-unique way on the tree, or we can think of the tree as imposing a condition on the group: that instead, we should look for a group that acts effectively. That is, we should impose the group identity

$$LR^{-1}L^2R^{-1}L = e \tag{3.1}$$

as a condition. In the next section, we shall see that this defines the modular group $PSL(2,\mathbf{Z})$. That is, we have that

$$F_{\{L,R\}}/\{LR^{-1}L^2R^{-1}L = e\} = PSL(2,\mathbf{Z})$$

FIGURE 3.2. A three-point turn



This underlying identity provides the connection between the modular group, which is well-known in number theory and the theory of elliptical equations, and the period-doubling fractals.

Before exploring the group structure more fully, its worth doing a quick calculation. Choosing the matrix representation

$$L = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

one can quickly verify that $LR^{-1}L^2R^{-1}L = -I$ where I is the identity matrix. It is the appearance of this minus sign that makes the resulting group $PSL(2,\mathbf{Z})$ and not something else: we have to identify the matrices $-I$ and $+I$ both with the group element e, if we are to get the symmetries of the unrooted infinite binary tree.

The unrooted infinite tree of figure 3.1 re-appears again in figure 8.1, where it is embedded into the hyperbolic plane. Section 8 explores the geometric and hyperbolic structures associated with it.

3.1. **Linear Groups.** Throughout the development above, various two by two matrices appeared and were employed in curious ways. It is now appropriate to review some of the properties of such matrices; they play many important roles in physics and mathematics, and so serve to tie the Minkowski question mark into these broader outlines.

The most general case is the *general linear group* $GL(2,\mathbb{C})$. This is the set of 2x2 matrices

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{3.2}$$

with entries $a,b,c,d \in \mathbb{C}$ from the complex numbers, whose determinant $ad-bc \neq 0$ is not zero. The non-vanishing determinant means that a given matrix may be inverted; as groups must always contain their inverse.

Thanks to the Möbius transformation 2.5, this group can be seen to act on the complex plane, taking a value $z \in \mathbb{C}$ to a value $(az+b)/(cz+d)$. In fact, it acts on the upper half-plane, as it can be shown that Möbius transformations always take complex numbers with a positive imaginary part to another complex number with positive imaginary part. The Möbius transformation introduces a bit of redundancy, however, as both A and $-A$ result

in the same transformation. Thus, it is common to introduce $PGL(2,\mathbb{C})$, the *projective linear group*. It is obtained by moding out by plus and minus the identity matrix I:

$$PGL(2,\mathbb{C}) = GL(2,\mathbb{C})/\{+I,-I\}$$

The cosets of $PGL(2,\mathbb{C})$ are $\{+A,-A\}$ of matrices together with their negative. The group $PGL(2,\mathbb{C})$ finds a remarkable application in the theory of 3-manifolds, as it is able to represent the set of conformal (angle-preserving) maps of a three-dimensional ball into three-dimensional space[6, 7].

The determinant is an overall scaling factor that can be scaled away, and so the *special linear group* $SL(2,\mathbb{C})$ consists of matrices with unit determinant. This again forms a group, since, for any two matrices A,B, one has $\det AB = \det A \det B$. The special linear group is remarkable for several distinct reasons. First, a map of the (whole) complex plane to (all of) itself is conformal if and only if it is an element of $SL(2,\mathbb{C})$[8]. In physics, specifically in special relativity, it describes the transformation properties of a spinor under special-relativistic changes of coordinates. The product of $SL(2,\mathbb{C})$ and its complex conjugate forms the adjoint representation of $SO(3,1)$, which is the Lorentz group, which describes the transformation of vectors in special relativity[9].

It is often convenient to work with the group $S^*L(2,\mathbb{C})$, which is defined as the group of 2x2 matrices whose determinant is +1 *or* -1. It was already noted above that the partial convergents of a continued fraction, when arranged into a 2x2 matrix, will have a determinant of +1 or -1. Finally, $PS^*L(2,\mathbb{C}) = S^*L(2,\mathbb{C})/\{+I,-I\}$ is the projective version. Since the determinant of both plus and minus the identity is +1, the projective case does not mix the signs of the determinants! This is an important yet sometimes confusing detail: in two dimensions, $\det -I = \det I = 1$.

Restricting the matrix entries to real numbers, one obtains $SL(2,\mathbb{R})$, the *special linear group over the reals*. The projective linear group over the reals $PSL(2,\mathbb{R})$ plays an important role in the theory of Riemann surfaces, as it is the group of all orientation-preserving isometries of the upper half-plane.

The remainder of this article concerns itself with $GL(2,\mathbb{Z})$, the group of invertible 2x2 matrices with integer coefficients, and its subgroups $SL(2,\mathbb{Z})$ and $PSL(2,\mathbb{Z})$. Note that $S^*L(2,\mathbb{Z}) = GL(2,\mathbb{Z})$, since, if a 2x2 matrix has integer entries and is invertible, it must necessarily have a determinant equal to +1 or -1. The subgroup $SL(2,\mathbb{Z})$ consists of those matrices that have unit determinant. The subgroup $PSL(2,\mathbb{Z})$ consists of those matrices that have unit determinant, and where we are free to ignore an overall sign.

3.2. **The Modular Group.** The projective linear group of two by two matrices with integer coefficients is $PSL(2,\mathbb{Z})$. This group plays an over-arching role in number theory, and has many fascinating properties[8, 10]. Its most important property is that it is the group of isomorphisms of the planar grid of parallelograms. Take, for example, the square grid of all points (m,n) for integers m,n. This lattice is generated by two vectors e_1 and e_2

$$e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

in that the linear combinations $v = me_1 + ne_2$ visit every lattice point. These two generators are not unique; so, for example, $(1,-2)$ and $(0,1)$ will also generate the lattice. The general case is invariant under transformation by any matrix $A \in SL(2,\mathbb{Z})$; that is, if the two vectors e_1 and e_2 generate the grid, then so do Ae_1 and Ae_2. Conversely, any pair of generators of the grid can be expressed as a pair of columns in a matrix in $SL(2,\mathbb{Z})$.

The connection to fractions follows from the fact that the matrix entries are all relatively prime to one another. Writing the matrix as 3.2, one has that the pair of integers (a,b) have no common factors, and so the fraction a/b is in lowest common terms, as are the fractions b/c, c/d and d/a.

The connection to the Minkowski Question Mark becomes visible when one tries to enumerate all of the possible elements of $PSL(2,\mathbb{Z})$. There are several ways in which this is commonly done. One way is to discover that $PSL(2,\mathbb{Z})$ is isomorphic to the free group $\mathbb{Z}_2 * \mathbb{Z}_3$. That is, one defines two matrices V,P as

$$V = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad P = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}$$

which have the property that $V^2 = P^3 = -I$. It may then be shown that every element $A \in SL(2,\mathbb{Z})$ may be written as

$$A = (-I)^r P^{p_0} V P^{p_1} V P^{p_3} V \cdots V P^{p_n} \tag{3.3}$$

for some finite integer n. Furthermore, it can be shown that this enumeration is unique[8], provided one sticks to the enumeration that r can be only zero or one, p_0 and p_n can take the values 0,1,2 and the remaining p_k can only take the values 1,2. By "ignoring" the value of r, one gets an element of $PSL(2,\mathbb{Z})$.

The enumeration can be accomplished by using elements of the dyadic rationals $\mathbb{Q}_2$ (eqn 1.3) to encode the values p_k. Explicitly, write an element of $\mathbb{Q}_2$ on the unit interval as

$$x = \sum_{k=1}^{n+1} b_k 2^{-k}$$

with $b_k \in \{0,1\}$. For each such x, set $p_k = b_{k+1} + 1$; this gives an element $x \mapsto A(x)$. Seven additional elements can be obtained by multiplying $A(x)$ by P on the left, the right, or multiplying by -1, thus allowing $SL(2,\mathbb{Z})$ to be enumerated by elements of $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Q}_2$. We shall call this the "dyadic enumeration".

A different pair of generators L,R (for "left" and "right") may be obtained by defining

$$L = V^{-1}P^2 = -VP^2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad R = V^{-1}P = -VP = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

The next section will show that these correspond to left and right moves on the binary tree, from which they earn their name. Using these definitions of L and R, the enumeration 3.3 can be brought into the unique form

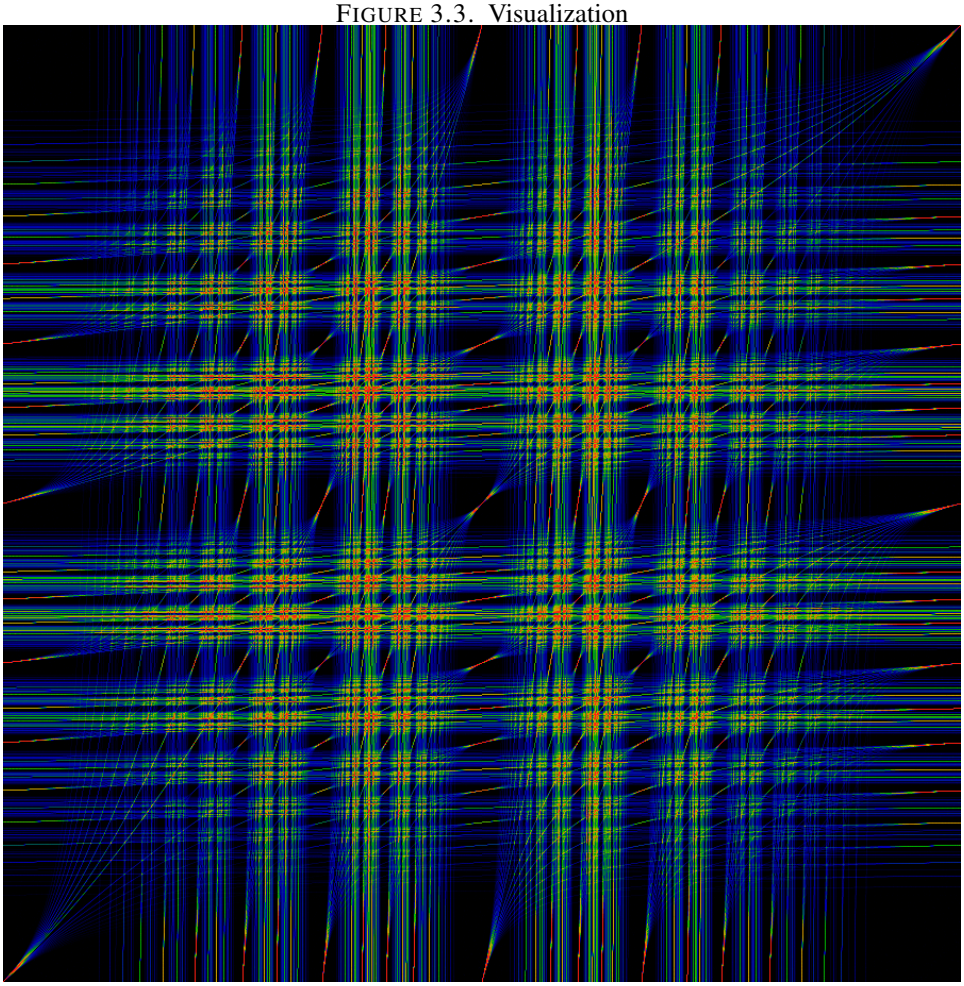
$$A = (-I)^s \left(R^{-1}L\right)^{p_0} L^{a_1-1} R^{a_2} L^{a_3} \cdots R^{a_n} \left(R^{-1}LR^{-1}\right)^{v_0} \tag{3.4}$$

for finite n. As before, this is a unique enumeration of the elements of $SL(2,\mathbb{Z})$. Here, s can be only zero or one. The value of p_0 is exactly as before, since $P = R^{-1}L$. There may or may not be a terminating $V = R^{-1}LR^{-1}$, so that v_0 is 0 or 1, depending on whether p_n was zero or not. The remaining coefficients a_k are all positive integers, except for a_n which may be zero, so as to describe the general case. Clearly, the integers a_k offer a run-length encoding of the number of times that p_j is one or two:

$$A = (-I)^{r+a_1+\cdots+a_n} P^{p_0} \underbrace{VP^2VP^2\cdots VP^2}_{a_1-1} \underbrace{VPVP\cdots VP}_{a_2} \underbrace{VP^2VP^2\cdots VP^2}_{a_3} \cdots \underbrace{VPVP\cdots VP}_{a_n} (V)^{v_0}$$

This form makes it clear that the elements of $SL(2,\mathbb{Z})$ can be enumerated by writing a rational number q as finite-length continued fraction. That is, let $q \in \mathbb{Q}$ on the unit interval

FIGURE 3.3. Visualization



This figure provides a visualization of $SL(2,\mathbb{Z})$ as Möbius transforms acting on the unit square. Specifically, every possible matrix A given by eqn 3.3 (with no initial P and with a trailing V) is generated, corresponding to all dyadics with denominator less than or equal to 2^{12}. Each such A is then treated as a Mobius transform $y = (ax+b)/(cx+d)$. A density scatterplot is then taken of the points $(x, \lfloor y \rfloor)$ with uniform distribution for x combined with a unform distribution for y. The brighter/redder the color, the greater the density. Note how the diagonals resemble those of figure 8.4. The same scatterplot, using elements starting with P looks the same, but with the diagonals missing. It should be clear that the intersection of a horizonatal or a vertical line with this plot will have the distribution shown in figure XXX (wtf, where is the measure? Where is this figure?)

---

be written as

$$q = [0; a_1, a_2, a_3, \cdots]$$

then the corresponding $q \mapsto A(q)$ belongs to $SL(2,\mathbb{Z})$. There are 11 additional elements that correspond to this q, since s and v_0 can take one of two values, and p_0 one of three.

Thus, the elements of $SL(2,\mathbb{Z})$ can be enumerated by $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Q}$. There are more enumerations possible; others will be given below.

The L,R form of the group elements begs the question, what happens if some of the a_k are negative? Although legitimate group elements result, the enumeration is no longer unique. In particular, $L^{-1} = -PV$ and $R^{-1} = -PV^2$ and so strings containing negative powers of L and R can be converted to strings containing only positive powers, after re-grouping terms.

The identity $P^3 = V^2 = -I$ with $P = R^{-1}L$ and $V = -L^{-1}RL^{-1}$ leads to the curious identity $LR^{-1}L = R^{-1}LR^{-1}$. This is reminiscent of the identities defining the braid group. The braid group describes what happens when one inter-twines a set of strands, that is, when one braids them. The braid group, in a certain sense, generalizes the permutation group, in that, instead of just permuting objects, one tracks their histories as well, by attaching world-line paths or strings to show where they came from. The general braid group B_n of n strands has $n-1$ generators b_k, which represent the right-handed exchange of a pair of neighboring strands. These obey the relations $b_k b_{k+1} b_k = b_{k+1} b_k b_{k+1}$, which show that there are two equivalent ways to exchange a neighboring set of three strands in the braid. There are no additional relations between the b_k. For the case of $n = 3$, the braid group B_3 has two generators, b_0 and b_1. Identifying $b_0 = L$ and $b_1 = R^{-1}$, it is clear that the braid group B_3 has a representation in terms of 2x2 matrices. As there are no further relations between L and R, then clearly, B_3 is isomorphic to $SL(2,\mathbb{Z})$[xxx provide ref].

Rather remarkably, the free group in two letters appears as a subgroup of $SL(2,\mathbb{Z})$! This can be most easily shown by re-interpreting the (free) monodromy group around two branch points as a subgroup of the braid group. It is well known that the set of homotopy loops that can be cast around two points forms the free group $\mathbb{Z} * \mathbb{Z}$ in two letters. That is, letting X denote a loop that winds around one point, and Y a loop around the other point, a general element of the free group is $X^k Y^m X^n Y^p \cdots$, where k,m,n and p are any, positive or negative, integers, with no further constraints on what values can appear. Fixing two strands of the braid group B_3 so that they are rigid and not movable, the third strand can be made to wind around the first two, as long as its tip always returns to its starting position. Clearly, the third strand can wind around in every which way, thus forming the free group. In terms of the group generators, these moves correspond to $X = b_0^2 = L^2$ and $Y = b_1^2 = R^{-2}$. Because only the squares appear, the previously established identity $LR^{-1}L = R^{-1}LR^{-1}$ cannot be applied to limit which combinations of X and Y are unique.

Another commonly used pair of generators are S,T with $S = V$ and $T = R$. These are of interest primarily because the Möbius transformations $S\tau = -1/\tau$ and $T\tau = \tau + 1$ give a particularly simple form to the invariance properties of modular forms[10]. The general group element $B \in SL(2,\mathbb{Z})$ may be written as

$$B = T^{m_1} S T^{m_2} S T^{m_3} S \cdots S T^{m_n} \tag{3.5}$$

for some integers m_k, which may be taken to be positive or negative. In this case, the representation is not unique, as again, negative powers may be converted to positive powers. Comparing to the left-right generators, one has $L = -TST$ and so

$$L^{a_1-1} R^{a_2} L^{a_3} \cdots R^{a_n} = (-I)^{a_1+a_3+\cdots-1}\, T \left(ST^2\right)^{a_1-1} T^{a_2} \left(ST^2\right)^{a_3} \cdots T^{a_n-1}$$

The requirement for the enumeration 3.4 to be unique was that all of the a_k be positive integers. The equivalent requirement for uniqueness on the ST enumeration then requires that, in general $m_k \geq 2$. An intuitive way for understanding this is to note that $P = ST$, and so there are string identities such as $STSTST = -I$.

FIGURE 4.1. p-adic Tree

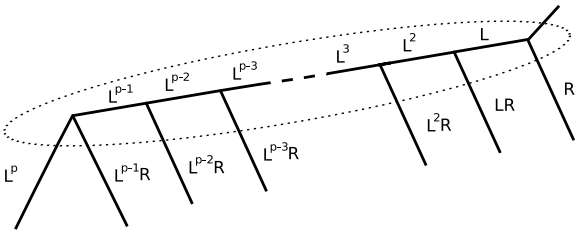
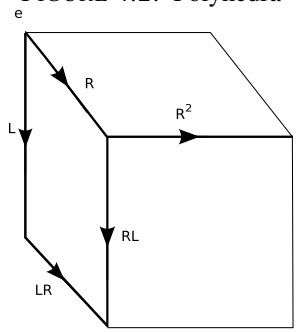


Figure above shows a tree with p branches, made from a binary tree by equating nodes.

FIGURE 4.2. Polyhedra



This figure shows a binary tree wrapped onto a polyhedron, a cube in this particular case. Note that this implies that its a group, since inverse elements are now clear. Note that unlike the p-adic tree, this has only pure torsion. The general case is a labeled graph, aka a semiautomaton.

Another case are walks on ZxZ, the square, flat euclidean grid. Here, we can take L to be walk up and R to walk right. Clearly this has all L's and R's commuting, so this is the fully commutative monoid, which extends to the free abelian group in two generators. The paritially commutative subsets are called "traces" or "history monoids" or "trace monoids" or "semi-commutative monoids", and correspond to systms of communicating finite state machines.

XXX Also mention symplectic group.

4. FREE MONOIDS

XXX ToDo, write this section. Discuss generalities first, then note all the things one can do with a binary tree.

- Define a free object.
- Define a free monoid.
- Define a presentation
- Define an act of monoid on set, aka semiautomaton.
- Discuss presentations with torsion, vs. those with that remain free.

Figures:

The general case is a semiautomaton. Consider a finite set of states (aka graph vertices). Consider state transitions driven by L,R; these are labeled arrows from one state to another. Starting from any given node, these show how to project the binary tree onto the graph. A special case of this are polyhedra. The general case are the regular languages.

Note that alternating algebras i.e. Grassmanians, can be mapped into a binary tree. This includes in particular the superalgebras.

Other possible embeddings of subtrees into trees.

Consider, for example, the binary tree as the set of all possible strings in L and R. Pick two arbitrary elements γ_1 and γ_2 from the monoid, and make the identification $L \mapsto \gamma_1$ and $R \mapsto \gamma_2$. Then the resulting set of strings in γ_1 and γ_2 is then a subtree of dyadic tree, and is, in particular, once again a dyadic tree. There are an infinite number of such subtrees, and they need not even be regular as the one above. For any given leftwards or rightwards move, one can pick arbitrary some element γ to stand in its place. Thus, the general infinite dyadic subtree can be represented by a countable set of monoid elements $\{\gamma_k | k \in \mathbb{N}\}$, with the identification that, given a string in L and R, the letter in the k'th position is replaced by γ_{2k+1} or γ_{2k+2}; the root of the subtree being given by γ_0.

XXX ToDo: finish writing out the rest of this section.

The polyhedra correspond to automorphic forms. See geometry section below.

The general subsets correspond to regular languages that is, are isomorphic to finite state machines. This has two corollaries: 1) the number of finite state machines are countable. 2) for every rational number, there is a finite state machine. There is an explicit coding– convert rational number to continued fraction. Convert continued fraction into binary rep. The repeated elements are the Kleene star elements of the regular language. An alternate way of thinking about this is that the Kleene-star elements correspond to the algebraic numbers (as per traditional Gauss continued fraction to algebraic number construction).

Also – many ODE/PDE numerical solvers can be understood as just small finite state machines acting on the 2-adic cantor-set representation of the real numbers.

Also – consider Simon Plouffe-type Ramanujan identities extended to automorphic forms, and thence to finite state machines. These come from theta-function type identities. Theta functions are just functions which are invariant for certain symmetries, and co-variant for others. They express a duality. between the torsion and torsion-free subgroups.

5. Representations of the Dyadic Monoid

The dyadic monoid, defined as the abstract set 2.8 of self-similarities, has several possible representations, in terms of matrices, as well as other objects. These are discussed in greater detail here.

5.1. **The Modular Representation.** The generators g, r of the self-similarities of the question mark, acting on the continued fractions, were given in equation 2.6 as

$$g_C = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad r_C = \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}$$

The subscript C is used to indicate that these are the transformations discussed in section 2.3. These matrices do *not* belong to $SL(2,\mathbb{Z})$, since the determinant of r_C is -1. They do belong to $S^*L(2,\mathbb{Z}) = GL(2,\mathbb{Z})$, the group of matrices with determinant +1 or -1. The

group $GL(2,\mathbb{Z})$ can be built from $SL(2,\mathbb{Z})$ simply by adjoining a matrix

$$N = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

which has the property that $N^2 = 1$ and $\det N = -1$. This matrix, when used as a commutator, has the remarkable property of inverting L and R, so that $NRN = R^{-1}$ and $NLN = L^{-1}$. Then, using $g_C = L$ and $r_C = -RN$, the general self-similarity transform 2.2 may be written as

$$\gamma_C = g_C^{a_1} r_C g_C^{a_2} r_C \cdots r_C g_C^{a_n} = \begin{cases} L^{a_1+1} R^{a_2} L^{a_3} \cdots R^{a_{n-1}} L^{a_n-1} & \text{for } n \text{ odd} \\ F R^{a_1+1} L^{a_2} \cdots R^{a_{n-1}} L^{a_n-1} & \text{for } n \text{ even} \end{cases}$$

That is, the general element is a sequence of left and right moves. Here, for n even, the letter F was introduced to stand for a "flip",

$$F = VN = g_C^{-1} r_C g_C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

in that it converts L's to R's when it commutes: $LF = FR$ and $RF = FL$. As a Möbius transform, $F : x = 1/x$, while, for continued fractions, $F : [a_1, a_2, \cdots] = [0, a_1, a_2, \cdots]$.

5.2. **The Dyadic Representation as Affine Transformations.** The action of g and r on the dyadic numbers, given in eqn 2.1, can also be expressed in terms of 2x2 matrices. However, quite unlike the modular representation, these matrices are not at all a subset of $GL(2,\mathbb{Z})$. The action is that of a linear affine transformation. A general affine transformation is a map

$$x \mapsto ax + b$$

Affine transformations can always be written as a matrix equation by taking x to be a vector, and bumping up the dimension of the vector by one. Thus, for example, one writes

$$\begin{bmatrix} 1 \\ x \end{bmatrix} \mapsto \begin{bmatrix} 1 & 0 \\ b & a \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \begin{bmatrix} 1 \\ ax+b \end{bmatrix}$$

The affine representation is then

$$\begin{bmatrix} 1 \\ \gamma_D(x) \end{bmatrix} = \gamma_D \cdot \begin{bmatrix} 1 \\ x \end{bmatrix} = g_D^{a_1} r_D g_D^{a_2} r_D \cdots r_D g_D^{a_N} \cdot \begin{bmatrix} 1 \\ x \end{bmatrix}$$

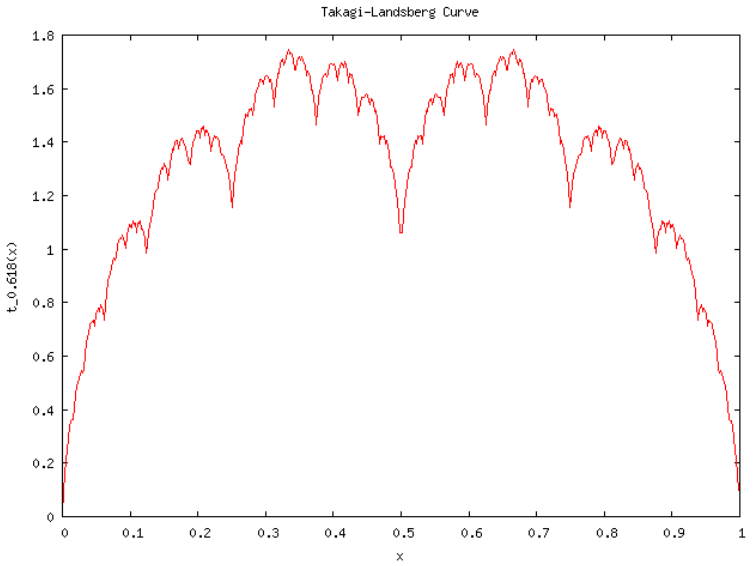
where

$$g_D = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \quad \text{and} \quad r_D = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \tag{5.1}$$

It is straightforward to verify that this gives the expression 2.9 for the action of a general monoid element on x. Although γ_D is thus representable as a two-by-two matrix, it is utterly different from γ_C. In particular, the determinant of $\gamma_D$is not 1. The γ_D are lower-triangular, and thus generate a "Borel monoid" , in analogy to the name "Borel group" given to the group of upper-triangular matrices.

5.3. **Higher Dimensional Affine Representations.** The dyadic representation can be generalized by using more general, $N \times N$ matrices in place of g and r. The full generalization is known as a *de Rham curve*, and is discussed in a later section, below. A particularly noteworthy example, however, is the three-dimensional representation, as it clearly embeds the dyadic representation. It occurs naturally in the description of the self symmetries of the Takagi curve.

FIGURE 5.1. The Takagi Curve



The Takagi-Landsberg or Blancmange Curve.

The Takagi-Landsberg curve is shown in figure 5.1. It is named after Teiji Takagi, who described it in 1901, proving that it was differentiable nowhere[11, 12]. The curve may be constructed as a superposition of triangle waves

$$t_w(x) = \sum_{n=0}^{\infty} w^n \tau^{n+1}(x) = \sum_{n=0}^{\infty} w^n \tau(2^n x) \tag{5.2}$$

where $\tau(x)$ is the triangle wave

$$\tau(x) = \begin{cases} 2x & \text{for } 0 \le x \le \frac{1}{2} \\ 2(1-x) & \text{for } \frac{1}{2} \le x \le 1 \end{cases}$$

Here, τ^k simply denotes the k-times composition of τ with itself: $\tau^k = \tau \circ \tau \circ \cdots \circ \tau$. The variable w is simply a parameter describing the curve; it may be taken as a real or complex value, but must have a magnitude of less than one in order for the series to converge. The figure 5.1 shows the curve for a parameter of $w = 0.618$.

Visually, the Takagi curve has a clear self-similarity. The action of g on t_w, understood to cut the interval in half, is to produce a scaled, sheared copy of itself:

$$[gt_w](x) = t_w\left(\frac{x}{2}\right) = x + wt_w(x)$$

By contrast, its left-right symmetry just means that $rt_w = t_w$. It is not hard to deduce that the action of a general element $\gamma = g^{a_1} r g^{a_2} r \ldots r g^{a_N}$ on t_w will be of the form $a + bx + ct_w$ for some constants a, b, c. This may be recognized immediately as a three-dimensional vector, and so the action of g and r can be given as operators acting on a three-dimensional space. Its also not hard to see that the action is linear, and so g and r are given by 3x3 matrices. Their explicit form may be given by making the identification of 1, x and t_w as the basis

vectors of the three dimensional space:

$$\begin{aligned} 1 &\longmapsto e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\ x &\longmapsto e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ t_w(x) &\longmapsto e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \end{aligned}$$

The transformation of each of these under g and r are given by

$$g_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 1 & w \end{bmatrix} \quad \text{and} \quad r_3 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Note that the upper-right 2x2 subparts of these matrices are just exactly g_D and r_D given above, in equation 5.1, which describe the action on the subspace spanned by $(e_1, e_2) = (1, x)$.

A similar curve, transforming as a four-dimensional representation, may be constructed from sections of a parabola; a five-dimensional representation can be built from sections of a cubic, and so on; this is explored in greater detail in [13]. The development begs several questions: what happens when g and r are replaced by general matrices, and even more generally, what happens when g and r are replaced by general maps? This is explored in the next section.

5.4. **Sequences.** The above can be extended to various integer sequences, such as the "fusc" sequence of the Calkin-Wilf tree – the fusc sequence gives the denominators of the Stern-Brocot tree. Likewise, the "batrachions":

- http://mathworld.wolfram.com/Hofstadter-Conway10000-DollarSequence.html
- http://mathworld.wolfram.com/MallowsSequence.html
- http://mathworld.wolfram.com/HofstadtersQ-Sequence.html

Also some discussion of Pompeiu functions that are self-similar.

5.5. **The Interval Map Representation.** Every element of the dyadic monoid corresponds uniquely to a node in the infinite binary tree. There are several ways to see this. First, the letters L and R may be taken to be the left and right moves through the binary tree; thus, any finite string consisting of the letters L and R uniquely specify a node on the tree. Alternately, g may be taken as a left-ward move on the tree, while r may be taken as a left-right reflection of the entire tree. Again, a finite-length string consisting of the letters g and r then uniquely identifies a node in the binary tree. It doesn't matter which navigation system is used; either can be converted to the other.

The subtree under a node on the tree represents an interval: the subtree encompasses everything from the left-most to the right-most sides of the subtree. Thus, the action of the dyadic monoid on the tree can be understood to be the action of a set of maps that map intervals to sub-intervals. Each interval is the range of a self-similarity map. To different interval maps were already given; these are 2.9 and 2.10.

The endpoints of an interval are the sup and inf of all the values on the corresponding subtree. For the Farey tree, the endpoints are always rational; for the dyadic tree, they are

dyadics. That they are rational is most easily demonstrated by noting that the interval is the range of an interval map, so the endpoints are given by $\gamma(0)$ and $\gamma(1)$. For the modular map 2.10, one endpoint is given by the rational number $\gamma_C(0) = [a_1+1, a_2, a_3, ..., a_{N-1}]$ (note the missing a_N). The other endpoint is $\gamma_C(1) = [a_1+1, a_2, a_3, ..., a_N]$ when $a_N \geq 1$, although it must be written as $\gamma_C(1) = [a_1+1, a_2, a_3, ..., a_{N-2}]$ when $a_N = 0$; this last special case is really just a reversal of the two endpoints. By considering interval endpoints, the question mark isomorphism 2.3 can be written in the shorter form

$$?(\gamma_C(x)) = ?(\gamma_C(0)) + \frac{(-1)^{N+1}}{(2^{a_1+a_2+...+a_N})} ?(x)$$

The endpoints of an interval are not completely arbitrary, but are correlated. Although one endpoint can be picked freely, so, one can pick an arbitrary $\gamma_C(0) = p/q \in \mathbb{Q}$, the choice for the other endpoint is limited: the other end of the interval is given by $a_N \in \mathbb{N}$ a positive integer. Since intervals correspond to elements of the dyadic monoid M, elements of M may be enumerated by $\mathbb{Q} \times \mathbb{N}$. This, of course, is hardly the only possible enumeration: of course, the intervals could have been enumerated by the node at the root of the tree; for the Farey tree, this node is a unique rational number, and so the intervals could be equally well enumerated just by $\mathbb{Q}$ alone. The difference between these two schemes is, of course, just Hilbert's infinite hotel.

The idea of self-similar maps that map intervals to sub-intervals is appealing because of its simple intuitive connection to many common fractals. When looking at the Koch snowflake curve, for example, one feels that one can point anywhere to find a self-similar copy: this is the freedom of choosing one endpoint to belong to $\mathbb{Q}$. But to find the entire run of the self-similar part, one's choices are far more limited: only certain strict sub-intervals appear; this is the more limited choice of $\mathbb{N}$ for the other endpoint. Thus, the $\mathbb{Q} \times \mathbb{N}$ enumeration of intervals is in this sense one of the more intuitive ways of specifying elements of the dyadic monoid.

5.6. **The Cantor Set.** The interval maps given by equations 2.9 or 2.10 can be generalized in several different ways. These maps had several properties:

(1) The maps $L(x)$ and $R(x)$ are maps from the unit interval to the unit interval.
(2) The maps $L(x)$ and $R(x)$ together are surjective onto the unit interval. That is, the union of the range of L and the range of R together cover the whole interval. There are no gaps.
(3) The range of L and R intersect at exactly one point: $L(1) = R(0)$.
(4) The maps L and R meet exactly in the middle; that is, $L(1) = R(0) = 1/2$.

One possible generalization is to relax the first condition, to let L and R be endomorphisms of some general space X. The result of this generalization is the de Rham curve[14], a continuous curve in the space X. The construction of this curve is briefly reviewed in the next section.

Another possibility is to relax the second condition (and so also the third and fourth). The result of doing so is the Cantor set. So, for example, the standard construction of the Cantor set may be given as follows. Let L and R be the maps

$$\begin{aligned} L(x) &= x/3 \\ R(x) &= (2+x)/3 \end{aligned}$$

so that L maps the closed unit interval $[0,1]$ to the closed interval $[0,1/3]$ and R maps it to $[2/3,1]$. Effectively, the open interval $(1/3,2/3)$ in the middle has been excluded. But this is just the first step of the standard construction of the Cantor set: the removal of the

middle third. This first step corresponds to the first level of the binary tree: there are two branches, and the middle has been excluded. Repeating this process, it should be clear that what remains, after an infinite number of steps, is the Cantor set. Furthermore, it should be clear that every remaining element of the Cantor set can be assigned a unique label, an infinite string of L's and R's applied in succession. The Cantor set can be visualized as the limit of the infinite binary tree: it is the set of "leaves" of the infinite binary tree.

All of the properties of the Cantor set are evident in the construction. The Cantor set is totally disconnected, in that, given one string of letters L, R specifying one path on the binary tree, every other string will take one to another point that is a finite distance away. The Cantor set is also a compact Hausdorff space. This may be argued by noting that the unit interval is a compact Hausdorff space, that both L and R are continuous maps, and that the continuous map of a Hausdorff space is a Hausdorff space. A totally disconnected, compact Hausdorff space is an example of a Stone space, a them that will be returned to when reconsidering the dyadic monoid as a semilattice.

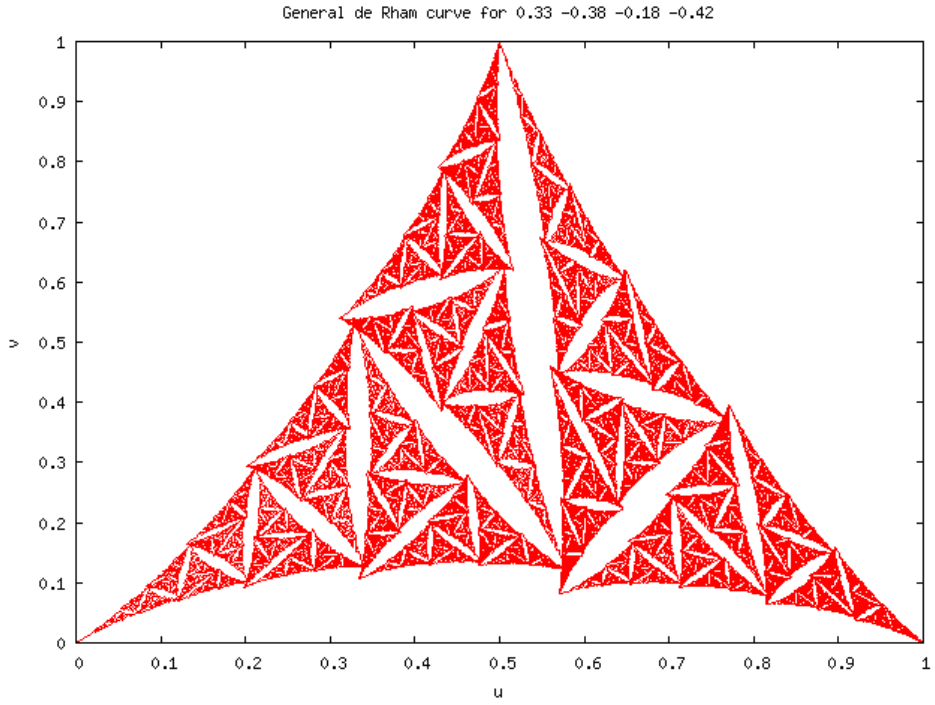
There are several noteworthy corollaries of this exercise. First is that the dyadic representation of a real number is isomorphic to the Cantor set, in that if $x = \sum_n b_n 2^{-n}$, the binary digits b_n can be taken to be the moves L and R on the binary tree. Next, the maps L and R can be quite general. One need not remove the middle third; one can remove less. One can arrange to remove even geometrically less at each step, so that the resulting set is no longer a Cantor set of measure zero, but the so-called "fat Cantor set", a Cantor set with finite measure. Notably, there are Cantor sets with measure one. In particular, if one removes the dyadic rationals from the real number line, what is left is a Cantor set, but it is a Cantor set with measure one (since the dyadic rationals are a set of measure zero). If instead one considers the Farey tree, it is then clear that the reals with all rationals removed is likewise the Cantor set. In essence, this is then a proof of something usually taken for granted: every irrational number x has a unique expansion $x = \sum_n b_n 2^{-n}$ in terms of binary digits b_n. It is the disconnectedness of the Cantor set that makes such an expansion unique.

5.7. **De Rham Curves.** The maps L and R may be taken to be endomorphisms of some general space X. Then the repeated iteration of of these maps, as described above, will result in the binary tree being mapped to a "dust" of points, called the "Cantor dust". With the right maps, the points of this dust can be joined together into a continuous line. If L and R are contracting maps, then, by the Banach fixed-point theorem, they will have fixed points. Let p_L denote a fixed point of L, so that $L(p_L) = p_L$, and likewise, let p_R denote a fixed point of R. Then, if p_L is in the basin of attraction R and p_R is in the basin of attraction for L, and furthermore, if the two maps are such that $L(p_R) = R(p_L)$, then the Cantor set is mapped into a continuous one-dimensional curve. This curve may be called a de Rham curve, in honor of Georges de Rham who described them and proved their continuity in 1957[14].

A "typical" de Rham curve is shown in figure 5.2. Its construction in terms of self-similar left and right pieces should be visually self-evident. It should also be clear that, with the right choice of parameters, a de Rham curve can be a space-filling curve. The set of all possible de Rham curves generated by two-dimensional linear affine transformations is five-dimensional (excluding degeneracies due to rotations, translations and rescaling), and is explored as a picture gallery in [15]. Several famous fractal curves fall in this class, including the Koch snowflake curve, the Peano curve and the Lévy C curve.

5.8. **Maximal Extension of the Dyadic Monoid.** The dyadic monoid M was constructed explicitly as the free product of the transformations g and r (or L and R), in which only

FIGURE 5.2. A Generic de Rham curve



non-negative powers were allowed to appear. This restriction implied that every $\gamma \in M$ corresponds to an interval map that is strictly shrinking (except for identity element); the unit interval is always mapped into a subinterval. Yet clearly, the interval map is invertible: for every $\gamma \in M$ there is a map $\gamma^{-1}(x)$ which maps some subinterval into the whole unit interval. The question then arises whether one can freely concatenate maps $\gamma(x)$ and $\gamma^{-1}(x)$ to form a group. The answer is no; this has already been touched on, and is explored more below.

Recall the definition of a *groupoid*: it is a set where every member has an inverse, and an associative product of two members is defined for some of the members (if the product were defined for all of the members, it would be a group, not a groupoid). For the collection of interval maps, every interval map is invertible, but not not every concatenation of interval maps is an interval map (because the domain of one may not intersect the range of the other). Thus, the collection of interval maps can be understood to be a *groupoid*; we call this the *dyadic groupoid*. It contains both the shrinking and the expanding maps.

Compare this to the definition of a *monoid*: it is a set where the associative product of every two members is defined, but some (or maybe all) of the members do not have inverses (if all members had inverses, it would be a group, not a monoid). If we consider the collection of interval maps, and consider only those which map larger intervals into smaller ones, then we have the *dyadic monoid*: The composition of any two maps is always defined, but none of the maps (except for the identity map) have inverses.

Thus, we see that the collection of interval maps forms a groupoid (the dyadic groupoid), and that a subset of these forms a monoid (the dyadic monoid).

Although the monoid M can be embedded in the group $GL(2,\mathbb{Z})$, not all of the group elements will have a well-defined action on intervals. One may then ask, what is the maximal extension $\widetilde{M}$ of M such that the elements of $\widetilde{M}$ can still be interpreted as maps of intervals? The rest of this section is devoted to exposing this extension; it will be shown that $\widetilde{M}$ is a groupoid, to be called the *dyadic groupoid*.

By considering the elements of M to be functions acting on the unit interval, we are, strictly speaking, talking about the action of M on the unit interval. To use precise language, this action is sometimes called the *set-theoretic representation* of M: it is a representation where elements are functions acting on a set (here, the set real numbers in the unit interval).

Every element $\gamma \in M$ defines an open interval υ_γ whose endpoints are given by $\gamma_C(0)$ and $\gamma_C(1)$. For the purposes of the following discussion, one can equally consider the intervals generated by $\gamma_C(x)$ or $\gamma_D(x)$; either will do, and so the subscript will be dropped. The length of this interval is strictly less than one: $|\upsilon_\gamma| = |\gamma(0) - \gamma(1)| < 1$ (except, of course, for the identity element, for which the interval is the unit interval). The set M consisted precisely of shrinking maps of the unit interval into itself.

For each $\gamma \in M$, the function $\gamma^{-1}(x)$ is well-defined on the domain $x \in \upsilon$ and has a range over the entire unit interval. This implies that $\forall \gamma_1, \gamma_2 \in M$, the map $(\gamma_1 \circ \gamma_2^{-1})(x)$ is well-defined on the domain $x \in \upsilon_2$, and has a range over υ_1. Thus, elements of the form $\gamma_1\gamma_2^{-1}$ also correspond to self-similarities, although they are not defined on the whole unit interval. Thus, one has that $\gamma_1\gamma_2^{-1} \in \widetilde{M}$, but, in general, $\gamma_1\gamma_2^{-1} \notin M$. That is, these elements belong to the groupoid but not, in general, to the monoid.

Theorem 1. *The only interval maps that provide a set-theoretic representation of the groupoid $\widetilde{M}$ are those maps of the form $\gamma = \alpha\beta^{-1}$ with $\alpha, \beta \in M$. There are no other maps. That is, the maximal extension of M as a set of interval maps is the set $\widetilde{M} = \{\gamma = \alpha\beta^{-1} | \alpha, \beta \in M\}$.*

Proof. Consider first the reverse ordering. That is, given $\gamma_2, \gamma_3 \in M$, construct the element $\gamma_2^{-1}\gamma_3 \in GL(2,\mathbb{Z})$, and ask if, or how, it might be a self-similarity. Let υ_2 and υ_3 be the intervals corresponding to γ_2 and γ_3, and let the superset relation $a \supset b$ simply denote that interval b is contained in interval a. If $\upsilon_2 \supset \upsilon_3$, then $(\gamma_2^{-1}\gamma_3)(x)$ has a domain of the entire unit interval, although its range is less than the whole interval. In fact, one has that $\gamma_2^{-1}\gamma_3 \in M$. This is easily demonstrated by appealing to the tree structure. The statement that $\upsilon_2 \supset \upsilon_3$ implies that υ_3 is represented by a sub-tree of υ_2. But the only way of getting from one node to a subnode is by navigating left and right branches till one reaches the desired subtree. The left and right descent operators are g and rgr, and so any descent path in the tree is given by some element $\delta \in M$. Thus $\gamma_3 = \gamma_2\delta$ or $\gamma_2^{-1}\gamma_3 = \delta \in M$ which completes the demonstration.

Consider next the case $\upsilon_3 \supset \upsilon_2$. The map $(\gamma_2^{-1}\gamma_3)(x)$ has a domain that is less than the entire unit interval, although the range is clearly the whole unit interval. One then has that $(\gamma_2^{-1}\gamma_3)^{-1} \in M$, and this may be shown using the same argument as in the last paragraph. The interval inclusion order implies that $\gamma_3^{-1}\gamma_2 \in M$, and since $\gamma_3^{-1}\gamma_2 = (\gamma_2^{-1}\gamma_3)^{-1}$, one is done. So again, the map induced by $\gamma_2^{-1}\gamma_3$ doesn't provide "anything new".

Finally, consider the case $\upsilon_2 \cap \upsilon_3 = \varnothing$. In this case, the map $(\gamma_2^{-1}\gamma_3)(x)$ is an invalid form: the range of $\gamma_3(x)$ does not intersect the domain of $\gamma_2^{-1}(x)$. Thus, the map $(\gamma_2^{-1}\gamma_3)(x)$ cannot generate a self-similarity. However, since $\gamma_2^{-1}\gamma_3 \notin M$ and $(\gamma_2^{-1}\gamma_3)^{-1} \notin M$, the element $\gamma_2^{-1}\gamma_3$ does provide "something new".

The net result of these considerations is that one does not gain anything new in considering a chain of elements of the form $\gamma = \gamma_1\gamma_2^{-1}\gamma_3\gamma_4^{-1}...$ with each $\gamma_k \in M$. A necessary condition that $\gamma(x)$ has a domain and a range that is not the empty set is that the intersection of intervals is not empty: $\upsilon_{2j} \cap \upsilon_{2j+1} \neq \varnothing$. But, as soon as this condition is imposed, each adjacent pair $\gamma_{2j}^{-1}\gamma_{2j+1}$ can be contracted with its neighbor on the left or right. Repeating this exercise, one finds eventually that either $\gamma(x)$ has a domain and a range that is the empty set, or that γ can be expressed as $\gamma = \alpha\beta^{-1}$ with $\alpha, \beta \in M$. This concludes the proof: the only interval maps in the interval map representation are those maps that belong to $\widetilde{M}$. □

Intuitively, the elements of $\widetilde{M}$ are those maps that map a subtree back up into the whole tree, and then map the whole tree back down to some other subtree. For every map that takes the whole tree to a subtree, the inverse map exists. These can be multiplied together, and, for many cases, maps in the forward and reverse directions can be composed together. However, the general composition of maps and their inverses does not exist, which is why $\widetilde{M}$ cannot be a full-fledged group. However, this intuitive description does show that $\widetilde{M}$ obeys all of the axioms of a groupoid, given below.

A *groupoid* is defined as a set G with an associative, partially defined multiplicative operator, and an involution that is the multiplicative inverse, which is defined for all elements. A "partially defined multiplicative operator" simply means that if $g, h \in G$, then their product $g * h$ may or may not be defined; if it is defined, then it belongs to G. An "associative partially defined multiplicative operator" means that, for $f, g, h \in G$, if $f * g$ is defined, and if $g * h$ is defined, then $(f * g) * h = f * (g * h)$. A groupoid also has a multiplicative inverse defined for every element, so that if $g \in G$, there exists a $g^{-1} \in G$ such that $g * g^{-1} = g^{-1} * g = e$. Here, e is the identity element; a groupoid by definition contains an identity element.

5.9. **Open topics.** A few unfinished thoughts:

What is the set of homomorphisms of $\widetilde{M}$? It should be a group, what is that group?

This groupoid is the fundamental groupoid of what space?

The general element of $\widetilde{M}$ is not defined on the whole unit interval, and the elements of $\widetilde{M}$ can be understood to for a sort-of sheaf or pre-sheaf. XXX. Expand on this idea.

6. THE DYADIC LATTICE

The introduction of the dyadic groupoid sidestepped some important questions: why can't the groupoid be extended to be a full group? Is the groupoid the most appropriate algebraic structure, or is there some other structure that more closely captures the structure of this thing? There is another, perhaps even a more appropriate, structure with which the set of self-symmetries can be understood. This is the lattice or Stone space[16] that is often used in general topology and is the focus of study in order theory.

6.1. **Inadequacy of the Dyadic Groupoid.** Calling the set of self-similarities a groupoid is inadequate, and fails to capture the full structure of the thing. There are several ways to see this. Corresponding to a move to sub-trees, there are inverse moves that take one back up the tree. If one moves to the left branch with L, there is an inverse move L^{-1}that takes one back to where one started. Similarly, there is an R^{-1} to undo a move R to the right branch. But L^{-1} and $R^{-1}$are the same thing: there is only one way to move back on the tree, and it does not depend on where one came from. Call it B, for "back", so that $B = L^{-1} = R^{-1}$, with the non-move given by $BL = LB = BR = RB = e$. A number of B's

can be concatenated together to indicate a number of upward moves. Thus, naively, one might want to extend the dyadic monoid by adjoining the letter B. Several uncomfortable problems arise with such an idea. First, the adjunction of B does not magically turn the monoid into a group: there is no unique element B^{-1} (it could be either L or R, of course). Next, the presence of B in any string of letters R,L automatically "erases" those letters: thus, there are no strings that mix together L,R and B. A third problem is that the operation of B is position-dependent. If one is at the the root of the tree, there is no way to back up any further. Thus, any given string B^n of n backwards moves may or may not be idempotent, depending on whether the starting point is at least n levels deep in the tree, or not. One concludes that there is a legitimate need to be able to formally discuss a back-up move, but that simply adding B to the monoid is insufficient.

There might be a temptation to try a different tack: to define a formal inverse. Every free monoid has a universal cover that is a group; the covering group is obtained by formally adjoining inverse elements to the monoid. The construction is straightforward, and is known as a "universal property", as it can be applied to the category of free monoids. In more concrete terms, this just simply means that one adds two elements R^{-1} and L^{-1} which are, by definition, inverses:$L^{-1}L = LL^{-1} = R^{-1}R = RR^{-1} = e$. However, in this universal extension, no further conditions are imposed; in particular, L^{-1} does not commute with R, and similarly R^{-1} does not commute with L. Thus, this formal extension does not match up with the idea of backwards moves on the binary tree.

If one considers the interval map representation, the formal extension also results in undefined moves. Consider, for example, $L(x)$ to be the function $L(x) = x/2$ that maps the whole unit interval to its left sub-half. Then $L^{-1}(x)$ is an inverse, whose domain is the half-interval [0,1/2]. Similarly, let $R(x) = (x+1)/2$ map the whole interval to the right half. What then should one make of $(L^{-1} \circ R)(x)$? The range of R simply does not intersect the domain of L^{-1}; so $L^{-1} \circ R$ is undefined as an interval map. The universal extension of the free monoid to a free group simply does not act on the binary tree in a meaningful way. One might try to rescue the situation by defining $L^{-1}(x) = 2x \bmod 1$, so that $L^{-1}(x)$ is well-defined on the entire unit interval. But then, $L^{-1}(x) = R^{-1}(x) = B(x)$, and one is back to having a "group element" that commutes, which is not what the universal extension did.

6.2. **Lattices and Semilattices.** A richer structure for the dyadic monoid can be found by observing that the thing that it acts on, the collection of intervals or infinite binary trees, has the structure of a semilattice on a partially ordered set. Actually, there are two possible semilattice structures; one, a distributive lattice inherited from the natural topology on the reals, and another, non-distributive semilattice that is more appropriate for working with the binary tree. But first, some definitions to anchor the topic are appropriate.

A *partial order* on a set A is a binary relation $\leq$ with three properties: 1) it is *reflexive*, so that $a \leq a$ for all $a \in A$; 2) it is *transitive*, so that if $a \leq b$ and $b \leq c$ then $a \leq c$, and; 3) it is *antisymmetric*, so that if $a \leq b$ and $b \leq a$, then $a = b$. A *partially ordered set*, or *poset*, is a set equipped with a partial order. The term "*partial order*" refers to the idea that not ever pair of elements in the set can be related with $\leq$; for some pairs, one just can't say. A *totally ordered set* is a poset for which the binary relation $\leq$ is defined between all pairs. Clearly, the set of all (complete, infinite) subtrees of the infinite binary tree is a poset, where $a \leq b$ means that a is a subtree of b. Here, as elsewhere in this article, a subtree is understood to be the entire infinite subtree anchored at a given node, rather than some other incomplete fraction.

Given any collection of subtrees, there is a unique smallest tree that contains the collection. This unique smallest containing tree, or least upper bound, is commonly called the

join of all the subtrees, and is denoted by $\bigvee$. The concept of a join is generally defined for any poset A. Given a subset S of the poset A, the join $a = \bigvee S$ is defined as an element $a \in A$ which is an upper bound to all elements in S, so that $a \geq s$ for all $s \in S$, but is also the least upper bound, so that, for any other upper bound b on S, one has $b \geq a$. If S is a two-element set $S = \{s,t\}$, one writes $s \vee t = \bigvee\{s,t\}$, the join of *s*and t, thus defining a binary operator $\vee$ between elements of A. It is not hard to see that all elements $a \in A$ are idempotent under the join operator: $a \vee a = a$, and that the join operator is symmetric $a \vee b = b \vee a$, and that the operation is associative: $a \vee (b \vee c) = (a \vee b) \vee c$ for all a,b,c.

For a general poset, joins need not always exist. Even when a subset has upper bounds, there may not be a least upper bound (such posets are called *directed sets*). However, when every finite subset of a poset does have a join, then one says that A has the structure of a *semilattice* or a *join-semilattice*. Clearly, the (poset of subtrees of the) binary tree is a join-semilattice, as any collection (even an infinite collection) of subtrees has a single unique join that contains them all. When a semilattice has joins even for infinite subsets, it is called *complete*; the binary tree thus forms a complete join semilattice.

Some authors also include the empty set as a possible subset, and define an the least upper bound of the empty set to be $0 = \bigvee \varnothing$. This new element 0 has the property that $0 \vee a = a \vee 0 = a$. In the context of the binary tree join-semilattice, there does not appear to be any need for this extra element. However, for the construction of the meet-semilattice from the binary tree, it will be needed.

The dual concept to the join is the meet, or greatest lower bound, which is obtained by reversing the direction of all of the inequalities in the definition above. Thus, the meet of a subset is $\bigwedge S$ and the meet of two elements is $a \wedge b$. Dual to the concept of zero is $1 = \bigwedge A$, which can be thought of as the greatest element in the poset. For the poset of subtrees, this element would be the entire tree. Meets on the poset of subtrees take one of two distinct forms. Given two subtrees, one is either a subtree of the other, or they are non-intersecting. Symbolically, either one has $a \leq b$, in which case $a \wedge b = a$ or neither a or b are a subtree of one another, in which case $a \wedge b = \varnothing$. This is a very different situation than that for the join on the binary tree: the join was always defined, for any pair a,b. By contrast, the meet of a pair of disjoint trees is not defined; it is the empty set. Thus, the (poset of subtrees of the) binary tree is, in this sense, not a meet-semilattice. This may be trivially fixed: adjoin the empty set $\varnothing$ to the poset of subtrees, so that the meet does become defined (it is the empty set!). This may seem trivial, but is belabored here, since it corresponds to a one-point compactification of the underlying structures that the binary tree represents. As was pointed out above, the removal of the rational numbers from the real number line results in a totally disconnected set, the Cantor set, which is modeled by the binary tree. Allowing something "in between" the gaps, even if that something is "the empty set", changes the structure of the "leaves" of the binary tree from being totally disconnected, to something connected. At this point, this may seem to be confusing word-play; it is only offered up as an introductory caution for the subsequent discussion of closure.

If a poset is both a join-semilattice, and a meet-semilattice, then it is called a lattice. Clearly, the poset of subtrees, adjoined with the empty set, is a lattice; it shall be called the "*dyadic lattice*".

XXX Also point out that there is a "Galois connection".

6.3. **Distributive and Modular Lattices.** The dyadic lattice has a number of properties; but foremost, perhaps, a few words on what it is not: it is not a Boolean lattice, it has no complements, it is not a distributive lattice, and it is not even a modular lattice. Lets look at

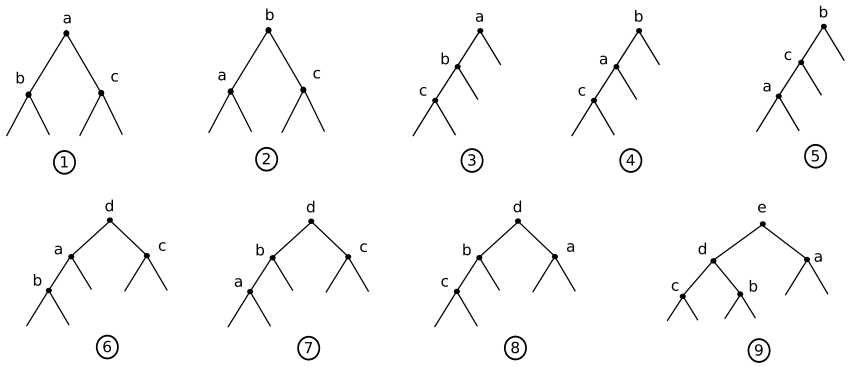
these in turn. A Boolean lattice is a distributive lattice with complements; since the dyadic lattice will have neither property, it won't be Boolean.

A *complement* on a lattice is defined as a unary operation $\neg : A \to A$ such that $\neg a$, the complement of a, obeys $\neg a \vee a = 1$ and $\neg a \wedge a = 0$. In the poset of subtrees, given some subtree a, there is clearly no other tree that could play the role of $\neg a$. One could construct a complement of sorts, but it would have to consist of many trees, rather than a single tree. This point will be explored in greater detail in a later section.

The distributive law is the identity $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$. If the distributive law holds for all a, b and c in a lattice, then the lattice is a *distributive lattice*. The theorem below shows that the dyadic lattice is not a distributive lattice.

Theorem 2. *The dyadic lattice is not a distributive lattice.*

Proof. To show this, one need only find a counter-example. This may be found by examining the various cases. There are nine distinct arrangements of the posets a, b, c, ignoring permutations of b and c. These are shown in the figure below.



The identity $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$ holds for all of the arrangements except for that in arrangement 6. To illustrate that example, consider first arrangement 1. There, one has $b \vee c = a$ since a is the smallest common tree that contains both tree b and tree c. Thus, clearly $a \wedge (b \vee c) = a$. On the other side, one has $a \wedge b = b$, since b is the largest tree that is contained inside of both a and b. Similarly, $a \wedge c = c$ and so $(a \wedge b) \vee (a \wedge c) = b \vee c = a$ and so the distributive law holds for arrangement 1.

For arrangement 6, one has $b \vee c = d$ and so $a \wedge (b \vee c) = a$. On the other side, $a \wedge b = b$ and $a \wedge c = \varnothing$, since tree a and tree c are disjoint. Thus, one has $(a \wedge b) \vee (a \wedge c) = b \neq a \wedge (b \vee c)$, and so the dyadic lattice does not obey the distributive law. □

Curiously, arrangement 6 is the only one that violates this distributive law. If instead, one considers the dual distributive law $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$, one finds that the dual is violated by arrangement 7, and by arrangement 9. The dual distributive law holds for all of the other arrangements.

A strictly weaker notion than distributivity is modularity. A lattice is said to be *modular* if $a \vee (b \wedge (a \vee c)) = (a \vee b) \wedge (a \vee c)$ holds for all a, b, c in the lattice. It is easily seen that every distributive lattice is a modular lattice.

Theorem 3. *The dyadic lattice is not modular.*

Proof. To verify modularity, one must examine 18 cases: the nine arrangements shown in the figure, and nine more with b and c reversed. Of the eighteen cases, only one does

not hold, which is the reversed version of arrangement 7. For this, the right-hand side is $a \vee (c \wedge (a \vee b)) = a \vee (c \wedge b) = a \vee \varnothing = a$ whereas the left-hand side is $(a \vee b) \wedge (a \vee c) = b \wedge d = b$. Thus, the dyadic lattice is not modular. □

6.4. **Möbius Function.** Given a finite partially ordered set, a Möbius function can be defined [xxx need reference]. The Möbius function is useful for inverting certain sums defined on posets. In some cases, it is straight-forward to provide a definition for some infinite posets as well. In particular, the Möbius function is well-defined on the infinite binary tree. It is given by:

$$\mu(a,b) = \begin{cases} -1 & \text{if } a \text{ is an immediate child of } b \\ +1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

Here, the expression "a is an immediate child of b" is the obvious: that $a < b$ and there is no x such that $a < x < b$. Here, the symbol $<$ refers to the obvious partial ordering on the binary tree; that is, $a < b$ if and only if a is in the subtree rooted at b. The Möbius function for this case can be obtained by considering finite binary trees; some simple but tedious computation will reveal the above answer.

6.5. **Ideals and Filters.** XXX make some intro commentary.

Define ideals and filters.

Finish writing me.

- Every element is an ideal, and in fact a principal ideal; this is trivial, since every element is represented by a single root of a tree.
- its meet-irreducible, therefore meet-prime. It has no join-primes
- i.e. every element is a prime ideal.
- Every ideal is the kernel of a semi-lattice homomorphism. The cosets of the resulting equivalence relation are the trees that are the direct parents of the ideal in question. Thus the quotient is always finite, and is always totally ordered.
- construct the complement via completed dual. i.e.take the dual, complete the dual, show its completion has complements!!!!
- The groupoid and its inverse together form a Galois connection.
- stone spaces

7. Compact Metric Spaces, Polynomials, Differential equations

Every compact metric space is the continuous image of the Cantor set[17]. This begs a host of (unanswered) questions: For which compact metric spaces is some remnant of the dyadic groupoid preserved? Or, perhaps more succinctly, for which compact metric spaces is the dyadic groupoid not evident? So, for example, the real numbers seem to have the structure of the Cantor set embedded in them, as exposed in this paper. How does this generalize to arbitrary compact metric spaces?

That is, the mapping of the Cantor set to the unit interval gives the unit interval a certain "hyperbolic" structure, as discussed in the section 8. This hyperbolic structure follows from the embedding of the dyadic groupoid (or binary tree) in the modular group, and essentially "explains" why iterated functions on the unit interval generate a fractal structure. So for example, many iterated functions on the unit interval are isomorphic to the Bernoulli map; this explains the fractal structure that results from iteration. How does this generalize to compact metric spaces in general?

The topological structure of a space is closely tied to the structure of the ring of functions on that space. If a compact metric space inherits a "hyperbolic" structure from its universal Cantor set covering, when and how is this manifested in the ring of functions? What subset(s) of this ring preserve the symmetries of the dyadic monoid?

To a large degree, it appears that polynomials "wipe out" or are incompatible with recursive, fractal structure: the smoothness of polynomials seems not to be able to support fractal self-similarity. There are a few exceptions: The parabola is a special case of the Takagi-Landsberg curve 5.2, with $w = 1/4$ (this special case, the construction of the parabola by midpoint displacement, being known to Archimedes[12]). Insofar as the Takagi curve is self-similar under a representation of the dyadic monoid, so is the parabola. Under what circumstances can other polynomial curves be considered to be self-similar, or in some way exhibit a (hidden) dyadic monoid symmetry? Another example shows itself on the complex plane: Modular forms have the symmetry of the full modular group, and thus of the dyadic monoid as well. What can be said about the set of complex-valued functions having only the symmetry of the dyadic monoid alone?

A related set of questions arise in the study of differential equations. Virtually all widely-studied differential equations have smooth, non-fractal solutions. Crudely speaking, the situation is analogous to a polynomial "wiping out" fractal structure. Yet, many important differential equations and/or integrable systems do exhibit a transition to chaos. So, for example, one can readily observe period-doubling in the laminar flow of the rising smoke from a cigarette, or in the oscillations of airflow over obstacles. This strongly suggests that the period doubling seen in differential equations is a signature of, an expression of the Cantor set in the space of solutions of these differential equations, a remnant of the universal nature of the Cantor set as a covering space for compact metric spaces. But what is the proper way of exhibiting this covering/embedding?

Note that the modular group is isomorphic to the symplectic group, that is, $SL(2,\mathbb{Z}) = Sp(2,\mathbb{Z})$. Insofar as the symplectic group is implicated in Hamiltonian dynamics, one may wonder if this can provide an opening for studying chaotic dynamics or integrable systems. The groupoid $\widetilde{M}$ of interval maps seems to exclude the symplectic case: the matrix $J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ does not belong to $\widetilde{M}$. Although one does have $\gamma_C^T J \gamma_C = \pm J$ for the Farey representation $\gamma_C \in \widetilde{M}$, one does not have a satisfying analogous relation for the γ_D. The closest that one comes to to find $K = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ which satisfies $\gamma_D^T K \gamma_D = K$, but this K only singles out the trivial representation. The dyadic and higher-order interval representations are not symplectic. However, there is clearly more to the story: after all, hyperbolic dynamics are chaotic, (mixing, and dissipative, even) and hyperbolic manifolds do have a Fuchsian symmetry.

8. Geometry and Hyperbolic Structure

The binary tree has a very nice, symmetric embedding into the hyperbolic plane, which will be explored in this section. The embedding is completely symmetric, in that, by marginalizing the importance of the root of the tree, one finds that the structure is completely symmetric and homogeneous. Formally speaking, the Cantor set has the property that, for any two points x and y, there exists a function f such that $f(x) = y$ and f is a homeomorphism. This is the definition of a homogeneous space, and so this makes the Cantor a homogeneous space. This set of homeomorphisms is given by the modular group $PSL(2,\mathbb{Z})$; the embedding makes this manifest. The action of the modular group on the

hyperbolic plane can then be reformulated as an action on the unit interval, by means of the usual embedding of the binary tree into the unit interval. These functions are sometimes called "hyperbolic rotations of the unit interval", and have some interesting properties of their own.

The hyperbolic plane is a two-dimensional surface, having a constant, uniform negative curvature of -1. As such, it is the hyperbolic partner to the usual to the usual two-dimensional sphere, having an everywhere constant curvature of +1, and the ordinary, flat Euclidean plane, having everywhere a curvature of 0. The hyperbolic plane can be conveniently represented as a subset of the complex plane (with a non-Euclidean metric) in two different ways: as the so-called "upper half-plane", and as the "Poincaré disk". The geometry of the hyperbolic plane is explored in a large number of classic texts[18, 19, 20][xxx need refs]; this section will assume some basic familiarity, and recap only a few basic definitions needed for the presentation.

The upper half-plane is given by the subset of the complex plane having positive imaginary values:

$$\mathbb{H} = \{x + iy = z \in \mathbb{C} \mid \Im z = y > 0\}$$

The hyperbolic metric on the upper-half-plane is given by

$$ds^2 = \frac{dx^2 + dy^2}{y^2}$$

This metric is often called the Poincaré metric. It is invariant under the action of the fractional linear transformations of $PSL(2,\mathbb{R})$. That is, suppose that z and z' are two different points in the upper half-plane, separated by some distance. Then, transforming with the fractional linear transform of eqn 2.5,

$$z \mapsto \frac{az+b}{cz+d} \quad \text{and} \quad z' \mapsto \frac{az'+b}{cz'+d}$$

leaves the distance between z and z' unchanged, whenever a, b, c and d are real-valued, and $ad - bc = 1$.

The upper half-plane may be mapped to the unit disk by means of the transformation

$$w = e^{i\phi} \frac{z - z_0}{z - \bar{z}_0}$$

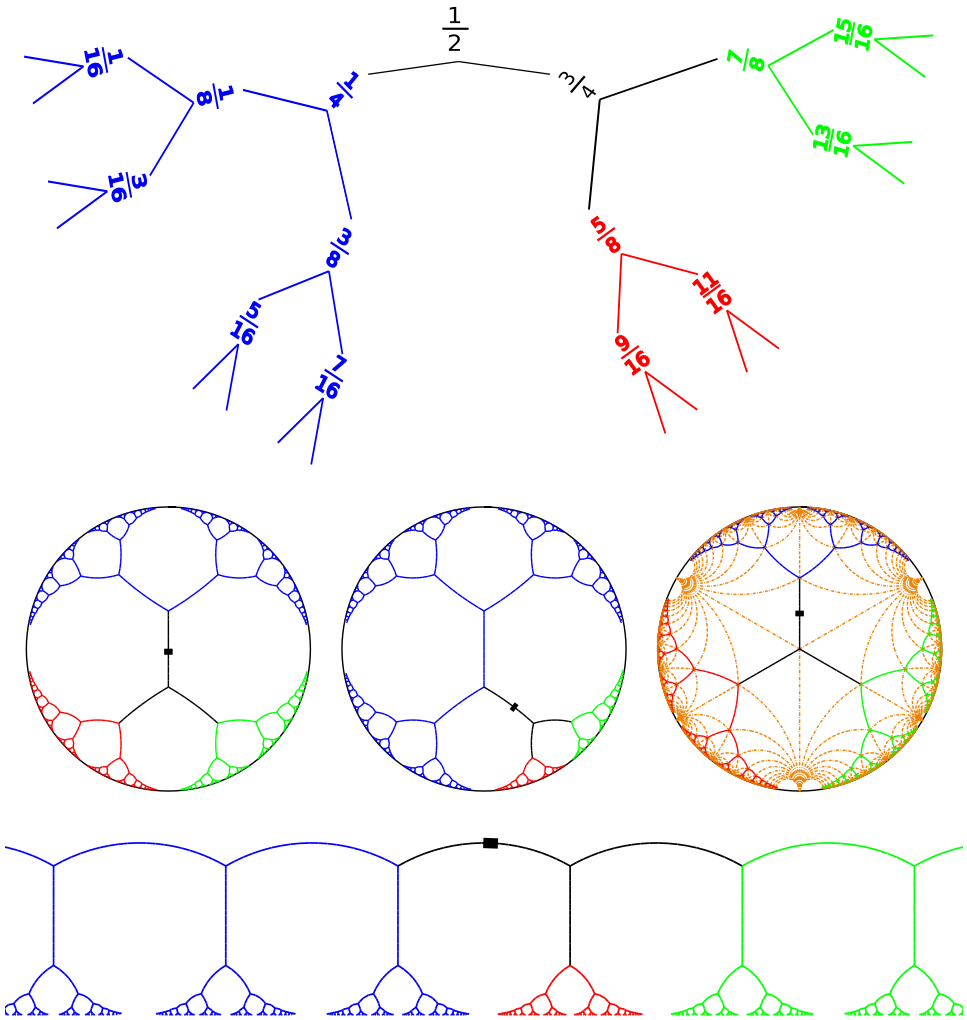
This mapping takes a point $z_0 \in \mathbb{H}$ and maps it to the center of the unit disk; the angular parameter ϕ just indicates an ambiguity with respect to the rotation of the disk. It is not hard to see that the whole upper half-plane $\mathbb{H}$ is mapped to the disk

$$\mathbb{D} = \left\{x + iy = z \in \mathbb{C} \mid |z|^2 = x^2 + y^2 < 1\right\}$$

The disk $\mathbb{D}$ is sometimes called the Poincaré disk.

A visual demonstration of the embedding is shown in the figure 8.1. The embedding proceeds in two steps. In the first step, one observes that by removing the root node of the binary tree, and simply replacing it by an arc, one obtains a very uniform graph, where all interior nodes always have three lines coming to them. The second step is to embed three points surrounding the origin, so that it will tile the hyperbolic plane under the action of $PSL(2,\mathbb{Z})$. The embedding that is illustrated puts vertices at $a = (1 + i\sqrt{3})/4$, $b = -1$ and $c = (1 + i\sqrt{3})/4$, connecting to the blue, red and green subtrees, respectively. The rest of the binary tree may be generated recursively, by repositioning each of these endpoints to the center of the disk, rotating the disk so that the incoming line segment extends from the center to $b = -1$, and then drawing two new line segments to a and c. This provides a geometric construction of the embedding.

FIGURE 8.1. The Embedded Tree



The rooted binary tree may be converted into the unrooted tree of figure 3.1 by unfolding it, and removing the root, as shown above. In the upper image, the root at 1/2 is removed and replaced with a single arc joining the left and right subtrees. This may be mapped to the Poincaré disk $\mathbb{D}$, as shown in the three middle images. The location of the original root is marked by a small black square. The third disk also shows the *fundamental domains* (see ahead to figure 8.6). The Poincaré disk may be mapped to the upper half-plane $\mathbb{H}$; one such mapping is shown in the bottom image. The small black square is located at $z = i$; the leaves of the tree run to the real axis $\Im z = 0$. The arced segments are circle arcs, so that, for instance, the arcs along the top are parts of circles of radius 1, centered at $z = n$. They meet at the intersections $z = n + \frac{1}{2} + i\sqrt{3}/2$. These figures make clear that the rooted binary tree, after removal of the root at at 1/2, is a Cayley tree (also called a Bethe lattice) with coordination number 3.

An algebraic construction of the embedding of the binary tree into the upper half-plane can be provided by using the generators L and R given previously:

$$L = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

Starting with the point $\rho = (1+i\sqrt{3})/2$ in the upper half-plane, one applies L to the the point $L\rho \in \mathbb{H}$, and then draws an arc from ρ to $L\rho$. This process is then repeated for R, and then recursively for all strings in L and R. The result is a binary tree, with a root at ρ, and running off to right of the half-plane. Because this tree is rooted at ρ, it does not fill the half-plane, but only the right quadrant. The other quadrant may be obtained by starting at the root $\rho - 1$, and recursively applying L^{-1} and R^{-1}. These two trees are then joined with a segment from $\rho - 1$ to ρ. The result is pictured in the bottom-most image of figure 8.1.

This embedding of the binary tree in the upper half-plane is invariant under the action of $PSL(2,\mathbb{Z})$. This may be easily seen from the algebraic construction: $PSL(2,\mathbb{Z})$ is generated by L and R (including the negative powers L^{-1} and R^{-1}, of course, as this is the group, not the monoid), whereas the binary tree consisted of arc segments connecting neighboring group elements, differing only by an L or R.

8.1. **Feigenbaum ratio.** Given a concrete mapping of the tree structure to the plane, it is worth asking: does the Feignbaum constant appear? The answer appears to be no, at least, not in the simplest way. Considering, for example, the imaginary coordinate of the vertexes in the upper half-plane, and take that to be the Feigenbaum bifurcation parameter. Taking the ratio of successive differences, one does not get the Feigenbaum constant; one gets, instead, a ratio tending towards 1.0, from above.

8.2. **Limit Points.** If the nodes of the binary tree are labeled, then the embedding of the binary tree into the hyperbolic upper half-plane, or the Poincaré disk, induces a mapping from tree coordinates to disk coordinates. Of particular interest are the coordinate mappings of the limit points, that is, of the "leaves" of the binary tree. There are a series of closely-related mappings; they will be needed in a later section and are thus spelled out here.

Consider first the mapping induced by taking the dyadic expansion of a real number in the unit interval, and replacing every occurrence of 0 by the matrix L and every occurrence of 1 by R. If the number is a dyadic rational, the expansion stops (ignoring the infinite trailing string of zeros). Suppose that the resulting matrix has matrix entries a,b,c,d, as in equation 2.5. The infinite trailing string of zeros corresponds to the matrix

$$L^{\omega} = \begin{bmatrix} 1 & 0 \\ \omega & 1 \end{bmatrix}$$

where $\omega = \infty$. Its not hard to see that the resulting product

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \omega & 1 \end{bmatrix} = \begin{bmatrix} a+b\omega & b \\ c+d\omega & d \end{bmatrix}$$

will map points $z \in \mathbb{C}$ of the complex plane to

$$z \mapsto \begin{bmatrix} a+b\omega & b \\ c+d\omega & d \end{bmatrix} : z = \frac{(a+b\omega)z+b}{(c+d\omega)z+d} = \frac{b}{d}$$

Thus, this procedure associates to every dyadic rational $0 \le x \le 1$ a positive real number b/d constructed from the dyadic expansion of x. Recognizing that this is a form of the de Rham curve, its clear that the function is continuous. Examining the construction process, it can be seen that the map is simply a pairing of the dyadic tree to the Stern-Brocot tree.

Let $\beta : [0,1] \to \mathbb{R}^+$ denote this mapping of the dyadic tree to the Stern-Brocot tree. Then, applying the analysis developed in previous sections, it is straightforward to determine that

$$\beta(x) = \begin{cases} ?^{-1}(2x) & \text{for } 0 \le x \le \frac{1}{2} \\ \frac{1}{?^{-1}(2(1-x))} & \text{for } \frac{1}{2} \le x \le 1 \end{cases}$$

This result is essentially a variant of the mapping between the Farey tree and the Stern-Brocot tree, given in eqn. 1.1. It is noteworthy, in that it provides a rapid algorithm to compute the inverse of the question mark function.

Consider next the embedding of the unrooted dyadic tree into to upper half-plane, as depicted in figure 8.1. This mapping takes the left half of the binary tree, and maps it into the left quadrant of the complex plane, and the right half of the tree to the right quadrant. If the dyadic tree is labeled as in figure 8.1, then one has a map $\delta : [0,1] \to \mathbb{R}$given by

$$\delta(x) = \begin{cases} \frac{-1}{?^{-1}(4x)} & \text{for } 0 \le x \le \frac{1}{4} \\ -?^{-1}(2-4x) & \text{for } \frac{1}{4} \le x \le \frac{1}{2} \\ ?^{-1}(4x-2) & \text{for } \frac{1}{2} \le x \le \frac{3}{4} \\ \frac{1}{?^{-1}(4-4x)} & \text{for } \frac{3}{4} \le x \le 1 \end{cases}$$

This unrooted dyadic tree may be embedded into the Poincaré disk using the mapping

$$w = \frac{z-i}{z+i}$$

which places $z = i$ at the center of the disk. Here, $z \in \mathbb{H}$ is a point in the upper-half-plane, and $w \in \mathbb{D}$ is the corresponding point in the Poincaré disk. This mapping wraps the unit interval around the perimeter of the disk in a counter-clockwise fashion, starting with $x = 0$ at the right-most edge of the circle. Again, this is illustrated in figure 8.1. The mapping is not linear. A little bit of work shows that it is given by

$$\theta(x) = \arctan \frac{-2\delta(x)}{[\delta(x)]^2 - 1}$$

A remarkable feature of this mapping is that it roughly impersonates the inverse of the question mark function. An even closer cognate is given by

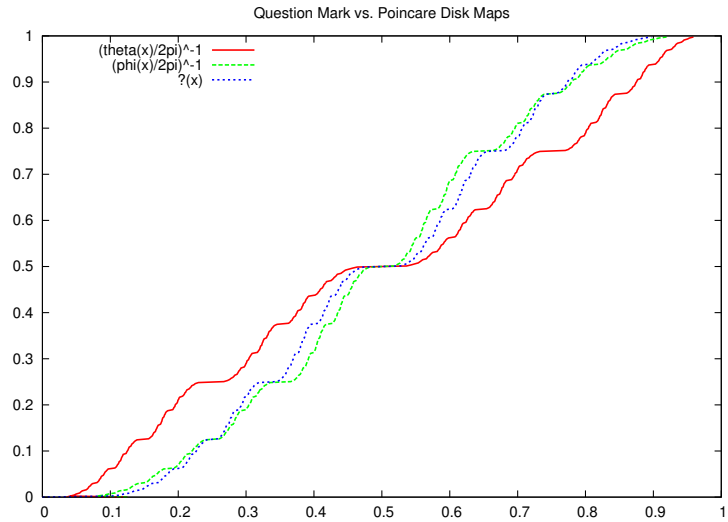
$$\phi(x) = \arctan \frac{-4\delta(x)}{[\delta(x)]^2 - 4}$$

which can be obtained as the projection of the real line to a circle. Both of these are shown in figure 8.2.

8.3. **Hyperbolic Rotations.** The geometric embedding of the (unlabeled) binary tree into the hyperbolic plane is invariant the action of $PSL(2,\mathbb{Z})$. However, if the nodes are labeled, then clearly, the action of $PSL(2,\mathbb{Z})$ permutes these labels. Insofar as the tips of the tree can be identified with the irrational numbers on the unit interval, so one has an action of $PSL(2,\mathbb{Z})$ on the unit interval. This action is briefly explored in this section.

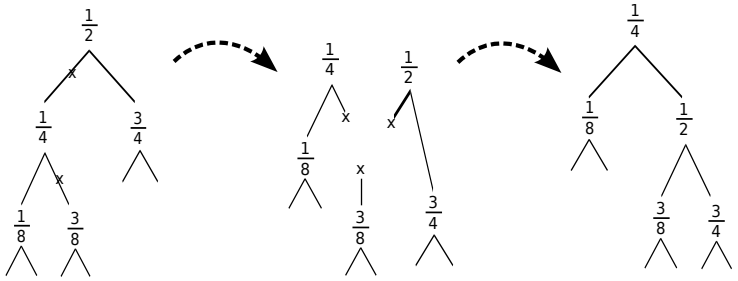
The action of R on the upper half-plane is to shift all nodes to the right, by one. Diagramatically, this change is shown in figure 8.3. This rotation preserves the arithmetic ordering of all of the labels on the tree; it is monotonically increasing. By simply comparing the tree before and after, it is clear that the action of this rotation on the unit interval is given by

FIGURE 8.2. Poincaré Disk Perimeter Functions



This figure shows a graph of the question mark function $?(x)$, and the inverses of the functions $\theta(x)/2\pi$ and $\phi(x)/2\pi$ discussed in the text. Note the similarity of the overall features, together with localized variations in slope.

FIGURE 8.3. Rotation of the Dyadic Tree



$$\rho_D(x) = \begin{cases} 2x & \text{for } 0 \le x \le \frac{1}{4} \\ x + \frac{1}{4} & \text{for } \frac{1}{4} \le x \le \frac{1}{2} \\ \frac{x+1}{2} & \text{for } \frac{1}{2} \le x \le 1 \end{cases}$$

The corresponding rotation for the Farey tree is given by the similarity transform $\rho_D \circ ? = ? \circ \rho_C$, with

$$\rho_C(x) = \begin{cases} \frac{x}{1-x} & \text{for } 0 \le x \le \frac{1}{3} \\ \frac{4x-1}{5x-1} & \text{for } \frac{1}{3} \le x \le \frac{1}{2} \\ \frac{1}{2-x} & \text{for } \frac{1}{2} \le x \le 1 \end{cases}$$

FIGURE 8.4. Rotations of the Unit Interval

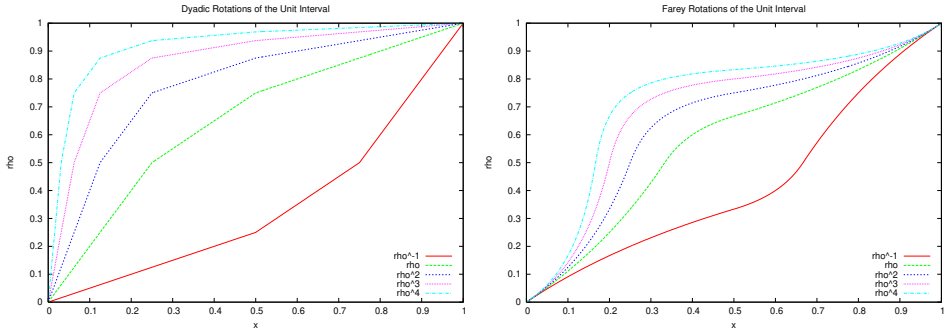
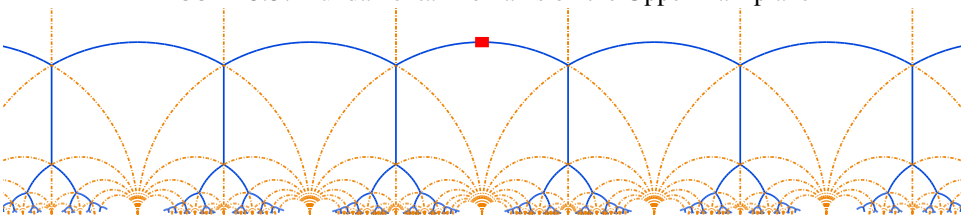


FIGURE 8.5. Fundamental Domains on the Upper Half-plane



This figure illustrates the traditional fundamental domains as used in number theory. Each is in the shape of a triangle, with two points in the upper-half plane, connected by a solid blue line, and a third point, the cusp, located at infinity, or, equivalently, on the real axis. The dashed yellow lines run from the interior points to the cusp. Note that the solid blue lines form the un-rooted binary tree, as demonstrated above.

To obtain this last form, one makes use of the identities given in section 2.4. It is perhaps curious that these rotations are linear and fractional-linear, respectively, and do not have a more complex form. Since both are monotonically increasing, both are invertible. A few of the compositions of these maps are shown in figure 8.4

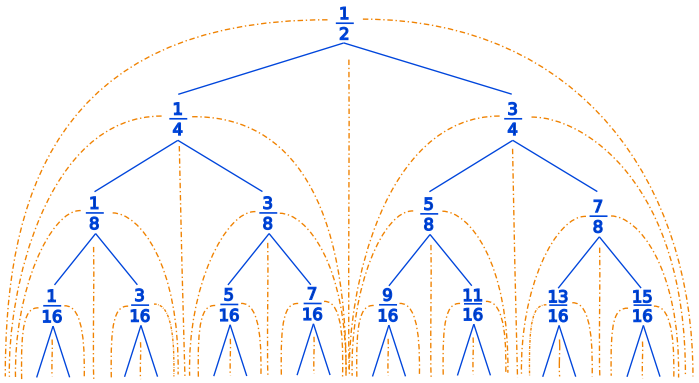
xxx finish me,

- review the non-order-preserving rotations.
- review rotations of sub-trees, (these are no longer rotations of the disk as a whole).

8.4. **Fundamental Domains.** Define a fundamental domain as the properly discontinuous action of a discrete group on a topological space.

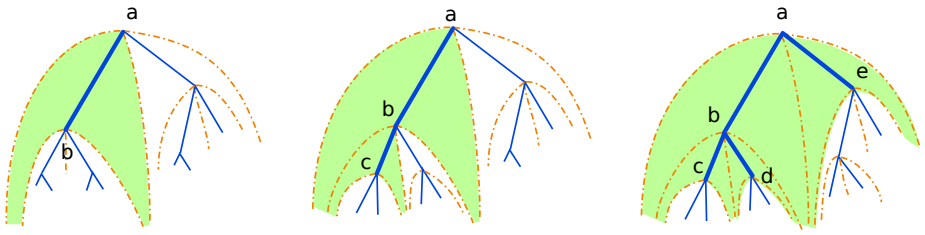
- Explain the figures 8.5,8.6 and 8.7 in greater detail.
- Talk about tiling the hyperbolic plane. Figure 8.7 shows tiling of the hyperbolic plane. Discuss how the "pumping lemma for regular languages"[21] is a form of tiling. Note that the tiling works not just in the forward direction, but also in the backwards direction; this allows the input to a finite state machine to extended to negative values (going back in time) similar to the way that the Grothendieck group is a construction that turns an abelian monoid into an abelian group.
- Figure 8.7 only shows some finite subtrees; should also show some infinite sub-trees.

FIGURE 8.6. Fundamental Domains on a Tree



A homotopically equivalent representation of fundamental domains, mapped to a binary tree. The arithmetic significance of the cusps becomes clearer when the geometry is deformed.

FIGURE 8.7. Domains from Subtrees



Consider a finite subtree of the infinite tree, shown in bold in the above figures. To the left and right of each segment lies a domain: these can be paired together, and, in a sense, "represent" that segment. These can be used to tile the hyperbolic plane. For example, the rightmost figure, consisting of four cusps, can be used to tile the plane.

Note also that such finite tree can be used to represent a finite state machine. For example, in the rightmost figure, the point "a" represents the initial state of the machine. The machine accepts strings in the two letters, "L" and "R". The input letter "R" takes the machine to state "e", which, if the tiling is to succeed, must be the same state as state "a". The input letter "L" takes one to state "b", while the input strings "LL" and "LR" take the machine to states "c" and "d", respectively, which, again, are exactly the same state as "a".

8.5. **ToDo:**

- Compare to L,R,B.
- classical automorphic forms on binary tree.
- automorphic forms for a finite state machine.
- moduli space of the automorphic forms corresponding to a given finite state machine; which are necessarily a subspace of the classical moduli space.

- eigenstates of the composition operator (of the rotations). (as separate chapter?)

9. XXXXXX

Everything below is misleadingly or confusingly stated and possibly wrong. Everything below this point needs to be corrected and re-written from scratch.

9.1. **More about representations of real numbers.** For one, the set $\mathbb{R}_D$ is larger than the set $\mathbb{R}$: the set $\mathbb{R}_D$ contains the elements 0.11111... and 1.0000... which are clearly distinct in $\mathbb{R}_D$ but represent the same number in $\mathbb{R}$. We have something similar happening in $\mathbb{R}_C$ in that a rational has the multiple representations $[a_0; a_1, a_2, ..., a_n, \infty, a_{n+2}, a_{n+3}, ...]$ where it does not matter what we pick for a_{n+2}, etc. Thus, both $\mathbb{R}_C$ and $\mathbb{R}_D$ are coverings of $\mathbb{R}$. So, for example, there exists a homomorphism, a projection, $\pi : \mathbb{R}_D \to \mathbb{R}$ such that $\pi(0.1111...) = \pi(1.000...)$, thus implying that Ker π is non-trivial. The real numbers are then defined as a quotient space of $\mathbb{R}_D$ by the equivalence relation induced by π.

Show $\mathbb{R}_D$ as being isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \cdots = \mathbb{Z}_2^\omega$but some care to be taken. Multiple possible topologies on $\mathbb{Z}_2^\omega$, including the product topology (which is what is wanted) and the box topology, (which is finer than the reals). Is there a need to discuss functions of a continuous variable, and so the dual spaces, the weak topology, etc?

Show that $\mathbb{R}_C$ is isomorphic to $\mathbb{Z} \times \mathbb{Z} \times \cdots = \mathbb{Z}^\omega$ and etc.

9.2. **The Modular Group, In General.** The close focus on the interval representation $\widetilde{M}$ in the above begs the questions "What about the other elements of the modular group? Where do they fit in?" These questions in fact have a simple answer: the modular group is the symmetry group of a two dimensional lattice. Let $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^2$ be two non-colinear vectors in the plane. Then the lattice $\Lambda(\vec{v}_1, \vec{v}_2) = \{p\vec{v}_1 + q\vec{v}_2 \ : \ p, q \in \mathbb{Z}\}$ generated by $\vec{v}_1, \vec{v}_2$ can be envisioned as a simple collection of parallelograms tiling the plane. The generators $\vec{v}_1, \vec{v}_2$ are not unique; in fact any other pair $\vec{w}_1 = a\vec{v}_1 + b\vec{v}_2$ and $\vec{w}_2 = c\vec{v}_1 + d\vec{v}_2$ will generate exactly the same lattice, if and only if $a, b, c, d \in \mathbb{Z}$ and $ad - bc = 1$. That is, $\Lambda(\vec{v}_1, \vec{v}_2) = \Lambda(\vec{w}_1, \vec{w}_2)$ iff

$$\begin{pmatrix} \vec{w}_1 \\ \vec{w}_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \vec{v}_1 \\ \vec{v}_2 \end{pmatrix}$$

where $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z})$. If we take $\vec{v}_1 = \vec{e}_1$ and $\vec{v}_2 = \vec{e}_2$ so that we have a square grid, then we can visualize the rationals as consisting of the points p/q in the upper-right quadrant of the grid, and specifically, those points visible from the origin, i.e. those points for which p/q is irreducible. The rationals confined to the unit interval correspond to those grid points lying above the horizontal axis, but below the 45°line extending to the upper right. Subintervals of the unit interval can now be understood to correspond tho those $(\vec{w}_1, \vec{w}_2)$ where both vectors are contained in this octant. The semigroup $M \subset SL(2,\mathbb{Z})$ of interval maps can now be understood as precisely those elements that stay within this octant. The "other" elements of $SL(2,\mathbb{Z})$ that are not in M are those that take intervals out of the octant. In particular, some elements of the modular group will take $(\vec{w}_1, \vec{w}_2)$ so that one lies in the left half-plane, the other in the right half-plane. Since the vertical line $p/q = 1/0$ corresponds to infinity, we see that such maps correspond to Möbius maps with a pole.

To summarize, the previous development of the interval representation seems to lend an air of mystery to the missing elements, those elements that are not in M. The mystery is dispelled in this wider picture, which accounts for all of the elements of the modular group.

Armed with this knowledge, we could seek to extended the definition of the question mark function to the entire real number line, as corresponding to the map between the full Farey tree (which contains all of the rationals) and the full dyadic tree, rather than just those pieces that correspond to the unit interval. The full Farey tree is shown in figure xxx (need figure here). However, there are several "natural" extensions of the dyadic tree to numbers larger than one, and so the appropriate extension of the question mark function is somewhat ambiguous.

One "natural" way to generate the dyadic tree is through the Takagi function recurrence relation

$$t_w\left(\frac{p}{2^n}\right) = w^{n-1} + \frac{1}{2}\left[t_w\left(\frac{p-1}{2^n}\right) + t_w\left(\frac{p+1}{2^n}\right)\right]$$

Besides generating the Blancmange curve, which we will explore in a later chapter, it also generates other interesting sequences if we pick a different set of starting conditions. For example, taking $w = 0$, and writing $t_0 \equiv d$, using the initial conditions $d(0) = 0$ and $d(1) = 1$, we generate the tree of dyadic numbers between 0 and 1; that is, we can promptly deduce that $d(x) = x$ for all $x = p/2^n$. If instead, we write $t_0 \equiv J$ with the boundary conditions $J(1) = 1$ and $J(1/2^n) = 2^n$, then we find that J generates the tree of dyadic numbers greater than one. This is shown in graph xxx (to-do Show the greater-than-one map explicitly.) Show also the extended Minkowski function for this dyadic sequence.

Another approach to the "problem" of the modular group being "bigger" than what is naively needed for fractal symmetry is to shrink its size. That is, one could try to construct a quotient group, mapping the eight octants into one. The benefit would seem to be that such a quotient construction would elevate the semigroup M to the status of a real group, containing inverses. However, it is not clear that any additional insight into fractal self-similarity is gained by doing so.

9.3. **Picard Group.** more about 3+1 "spacetime" generated by the complex numbers, (xxx this is actually called the Picard group see Fricke and Klein, circa 1897.) which is generated by $SL(2, Z \times Z)$ which is a subgroup of $GL(2, C)$. Here $Z \times Z$ are the Gaussian integers.

10. NOTES.

There is also a well-known representation of the Stern-Brocot Tree on the lattice $\mathbb{Z}\times\mathbb{Z}$, where a fraction p/q is denoted as the ordered pair (p,q). In this representation, the left-right navigation operators L and R have the values $L = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ and $R = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and thus positions are given as elements of $SL(2,\mathbb{Z})$. Maybe we'll elaborate on this later xxx. To do... elaborate.

To summarize, it is this expansion in binary digits that provides the underlying connection between period-doubling maps, such as the Mandelbrot Set, and Farey Numbers. Binary expansions, or code-words, occur naturally in the analysis of Duoady-Hubbard landing rays. We'll demonstrate an explicit mapping in a later section. XXX cut/reword this last paragraph.

10.1. **Some Curious Properties of the Question Mark.** If there is a 3-adic or p-adic generalization of the Minkowski Question Mark, it is not obvious; one 'obvious' generalization is

$$\sum_{k=1} (-1)^k \; 3^{-(a_1+a_2+...+a_k)}$$

but its highly discontinuous. Other generalizations based on roots of unity in the complex plane also don't seem to work. One might be able to get traction by looking at groups that have $SL(2,\mathbb{Z})$ as a subgroup but also have some p-fold symmetry.

11. Conclusions

To conclude, we've demonstrated two different binary trees commonly used in the representation of the real numbers, and have shown that the Minkowski Question Mark function is the mapping between these two trees. We've then reviewed the the modular group in terms of its action on trees, and showed that the self-similarity of the trees induces a fractal self-similarity on tree homomorphisms; in this case the homomorphism being the Question Mark function. We will explore the generalization of these ideas in the next chapter.

The wide-spread occurrence of the rationals fairly screams for an adelic (or p-adic) treatment of the subject matter. That is, in the above, we made no appeals to the closure of the rationals $\mathbb{Q}$ by $\mathbb{R}$ or by $\mathbb{Q}_p$; on the other hand, the interval representation makes numerous appeals to a total ordering. It would be interesting to see if and how any of the above conclusions are modified for the p-adic numbers.

Finally, we note that the interval representation is a topology, and that it is not exactly a trivial topology for a subset of the modular group. A more precise statement of the topological nature of the entire modular group, and how it relates to the interval representation, is called for.

11.1. **Handwaving insights.** Note that by imposing the modular group symmetry on the real number line, we've essentially introduced a hyperbolic manifold that is homomorphic to the real-number line. The existence of this hyperbolic manifold and its negative curvature essentially 'explains' why trajectories of iterated functions have positive Lyapunov exponents. Of course they do, since their 'true' trajectories should be considered to live on the hyperbolic manifold rather than on the real-number line.

References

[1] Alexander Bogomolny, "Stern Brocot Trees", , 1996-2006, http://www.cut-the-knot.org/blue/Stern.shtml.

[2] G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, 1938.

[3] J.H. Conway, *On Numbers and Games*, vol. 6 of *L.M.S. Monographs*, Academic Press, London, New York, 1976, iSBN 0-12-186350-6.

[4] A. Ya. Khinchin, *Continued Fractions*, Dover Publications, (reproduction of 1964 english translation of the original 1935 russian edition) edn., 1997.

[5] Linas Vepstas, "On the Beta Transform", *ArXiv*, 1812.10593, 2018, URL `http://arxiv.org/abs/1812.10593`.

[6] Bernard Maskit, *Kleinian Groups*, Springer-Verlag, 1988, iSBN 0-387-17746-9.

[7] Katsuhiko Matsuzaki and Masahiko Taniguchi, *Hyberbolic Manifolds and Kleinian Groups*, Clarendon Press, Oxford, 1998, iSBN 0-19-850062-9.

[8] Robert A. Rankin, *Modular Forms and Functions*, Cambridge University Press, 1977.

[9] Moshe Carmeli, *Group Theory and General Relativity*, McGraw-Hill, 1977, iSBN 0-07-009986-3.

[10] Tom M. Apostol, *Modular Functions and Dirichlet Series in Number Theory*, Springer, 2nd edn., 1990.

[11] Teiji Takagi, "A Simple Example of a Continuous Function without Derivative", *Proc Phys Math Japan*, 1, 1903, pp. 176–177.

[12] Benoit Mandelbrot, "Fractal landscapes without creases and with rivers", in *The Science of Fractal Images*, edited by Dietmar Saupe Heinz-Otto Peitgen, Springer-Verlag, 1988, p. 246.

[13] Linas Vepstas, "Symmetries of Period-Doubling Maps", , 2004, URL `https://www.linas.org/math/chap-takagi.pdf`, self-published on personal website.

[14] Georges de Rham, "On Some Curves Defined by Functional Equations (1957)", in *Classics on Fractals*, edited by Gerald A. Edgar, Addison-Wesley, 1993, pp. 285–298.
[15] Linas Vepstas, "A Gallery of de Rham Curves", , 2006, http://www.linas.org/math/de_Rham.pdf.
[16] Peter T. Johnstone, *Stone spaces*, Cambridge University Press, 1982, iSBN 0-521-23893-5.
[17] N.L. Carothers, *A Short Course on Banach Space Theory*, vol. 64 of *Student Texts*, London Mathematical Society, 2005, iSBN 0-521-60372-2.
[18] Hershel M. Farkas and Irwin Kra, *Riemann Surfaces*, Springer-Verlag, 1980.
[19] Jurgen Jost, *Compact Riemann Surfaces*, Springer-Verlag, 2002.
[20] Svetlana Katok, *Fuchsian Groups*, University of Chicago Press, 1992.
[21] John E. Hopcroft and Jeffrey D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley Publishing, 1979, iSBN 0-201-02988-X.

<LINASVEPSTAS@GMAIL.COM>