CUTTING SEQUENCES IN EUCLIDEAN AND HYPERBOLIC GEOMETRY

MERU BHANOT

Abstract. The most standard way of describing a geodesic (straight line) in the Euclidean plane is by its slope. In this paper, we present a symbolic way of encoding such a geodesic called the *cutting sequence*. We then prove that the cutting sequence and the continued fraction of the slope of a line contain the same information. We will then use the notions generated by this exploration to explain the associated topic of cutting sequences generated by geodesics in the Upper Half Plane model of Hyperbolic space.

Contents

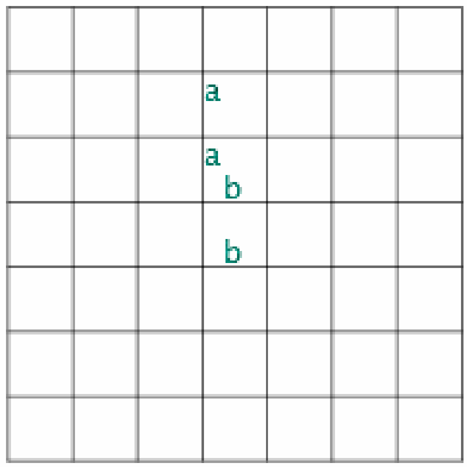
1. Cutting Sequences in Euclidean Space

Generally speaking, a *cutting sequence* is a series of markers that represent the path of a curve. In creating useful markers and studying the patterns that curves generate when they cross through them, we are able to determine information about the curves that was not initially clear, such as the slope of an arbitrary line. To explain how this general concept of "markers" can lead to a relevant mathematical result, we will begin with the simplest possible example.

Date: September 21, 2010.

1.1. **Lines on a Square Grid.** We will begin by taking a square grid and marking off each vertical segment with the letter a and each horizontal segment with the letter b, as shown below.

Figure 1



Note that since every vertical and horizontal segment is identical, we have, with these two classifications, labeled every portion of the grid.

Definition 1.1. The cutting sequence of a line C is defined as the sequence of sides a and b that one encounters when walking along C in a given direction from a given starting point.

For example, in Figure 2, we see the line C cuts through *abbabbabbab* when we start at the lowest point and travel upwards along the line. Since it is possible that a line will pass through a vertex of the grid, we will label such intersections as ab. While it is also possible to label these as ba, it is irrelevant in the long run, so long as we are consistent with our choice at each intersection of the line with a vertex.

1.2. **First Property of Cutting Sequences.** We will begin with the assumption that each square in our grid has length and width of 1. We will further assume that for any given line C, for every horizontal distance 1 the line travels, it travels λ vertically. Thus, the slope of the line $C = \lambda/1 = \lambda$.

Now let us consider the simplest case in which $\lambda = 1$. When we plot this on our grid, we see that every time the line goes through a vertical segment a, it must go through a horizontal segment b and then repeat the process. Thus the sequence we obtain is *abababab*...

We know, however, that this will not generally be the case. We will first consider the case in which the slope μ of a new line D must be such that $\mu > \lambda = 1$. (see Figure 3). What exactly would this fact imply about line $D's$ cutting sequence?

Recall that we began with the sequence *abababab*... for line C. Our new line D is more vertical, meaning that it intersects horizontal segments b more often than line C. This leads to the first property of cutting sequences.

FIGURE 2

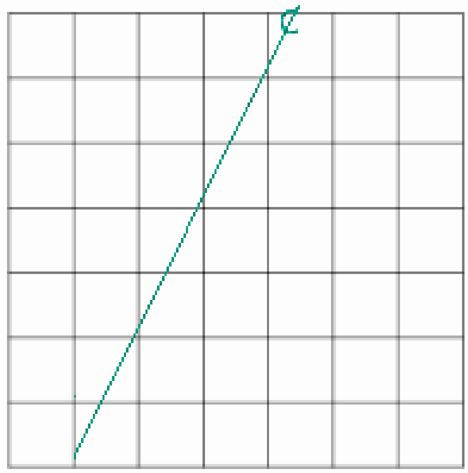
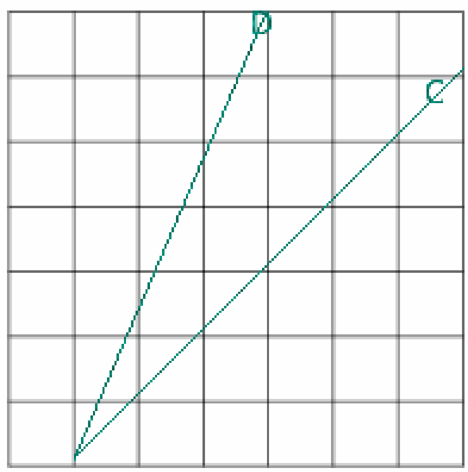


FIGURE 3



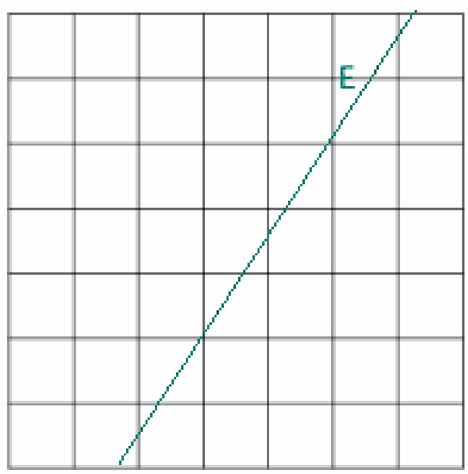
Property 1.2. *It must be the case that no two a's are ever adjacent to one another in D's cutting sequence if the slope of $D > 1$. In other words, the appearances of a in this case are* isolated. *Similarly, if the slope of $D < 1$, then the appearances of b will be* isolated.

Proof. Take for given that we have a line D with $slope(D) > 1$. Assume that in the cutting sequence of D there are two adjacent a's. This implies that the line D crosses through two vertical segments a without crossing through a horizontal segment b. If we isolate a square of the grid in which this occurs, and draw a diagonal from the bottom left to the top right, we get a segment of a line of slope 1 that crosses alternately through a and b. Note that line D, since it crosses through a twice without an b, must be more flat than the line of slope 1 that we constructed.

However, it was given that $slope(D) > 1$.
$\Rightarrow\Leftarrow$ □

1.3. **Second Property of Cutting Sequences.** Let us now consider a very specific line E of slope 1.5. Since $slope(E) > 1$, we know by Theorem 1.2 that the a's in $E's$ cutting sequence must be isolated, but do we also know something about the b's? Consider a portion of the line E that has just passed through a vertical segment a. If it passes through this segment in the bottom half, it will pass through one b and then pass back through an a. However, if it passed through the first a segment in the upper half, it will pass through two b's before passing back through an a. Since these are the only two cases, we see that a line of slope 1.5 has a cutting sequence that consists of a combination of ab and abb portions. In Figure 4, this sequence would be $abbababbaba$. (From now on, we will refer to abb as equivalent to ab^2.)

Figure 4



When we generalize this statement to lines of other slopes, we obtain the second property of cutting sequences:

Property 1.3. *If $slope(E) = \lambda > 1$, then between any two a's there are $[\lambda]$ or $[\lambda + 1]$ b's, where $[\lambda]$ is the integer part of λ. Similarly, if $slope(E) = \lambda < 1$, then between any two b's there are $[1/\lambda]$ or $[1/\lambda + 1]$ a's.*

Note that with all of these properties we are assuming that the lines we are considering have positive slope. This is not, however, a necessary assumption, but merely one we are making for the sake of simplicity. With some basic tweaking of the definitions we could also include lines of negative slopes.

1.4. **Combining the Properties.** Now that we have introduced these two properties, we will continue by using them in our originally stated goal: using cutting sequences to calculate the slopes of lines. To begin, we will introduce a few definitions that generalize our study:

Definition 1.4. A sequence that consists of a's and b's and that satisfies Properties 1 and 2 (Property 1.2 and Property 1.3 respectively) is called *almost constant*.

Note that this means that the cutting sequences of lines are "almost constant" sequences.

Definition 1.5. Futhermore, we say any almost constant sequence has a *value* $[\lambda]$ if $\lambda > 1$, or $[1/\lambda]$ if $\lambda < 1$.

For example, the sequence $ab^2ab^3ab^2ab^3\dots$ has value 2, because 2 is the smallest exponent of b.

We will now move away from expressly considering cutting sequences towards the study of these generic almost constant sequences. We will begin with the simple notion that most of these sequences, on the whole, will look something like this:

$$S = ab^n ab^n ab^{n+1} ab^n \dots$$

Of course these sequences could have a's and b's switched, and could vary differently in terms of the sequence of exponents and how they alternate, but the general structure will be the same. This allows us to simplify these almost constant sequences into what are called *derived* sequences.

Definition 1.6. Given an almost constant sequence S of value n, let $a' = ab^n$ and $b' = b$. Then there exists a *derived* sequence S' which can be created by substituting a' and b' elements into sequence S. If there are more a's than b's in the sequence, we define $b' = ba^n$ and $a' = a$ as expected.

For example, if we were to take the sequence:

$$S = ab^2ab^3ab^2ab^3\dots$$

the derived sequence would look like:

$$S' = (a')^2b'(a')^2b'\dots$$

because $a' = ab^2$ and $b' = b$.

Returning to our notion of the cutting sequences of lines, we see that cutting sequences can be infinite, meaning that it would be able to derive them an infinite number of times. This gives us another simple definition.

Definition 1.7. An almost constant sequence is called *characteristic* if after deriving the sequence any number of times it remains almost constant. In other words, if $s', s'', s'''\dots$ all exist, then the sequence is *characteristic*.

If we consider this for a moment, we see an interesting result. Given an almost constant sequence A that is characteristic, we are able to derive the sequence multiple times, and in the process we receive a new value for each derivation. Note that with this claim we are assuming that each derived series is almost constant, which will be explained in the following section.

Now that we have this infinite series of values, it is our eventual goal to prove the following theorem:

Theorem 1.8. *Given a cutting sequence of a line A with value v_0, such that v_1 is the value of its derived sequence, v_2 is the value of the second derivation, and so* *on,*

$$slope(A) = [v_0, v_1, v_2 \ldots] = v_0 + \cfrac{1}{v_1 + \cfrac{1}{v_2 + \cfrac{1}{v_3 + \cdots}}}$$

1.5. **Derived Sequences are Cutting Sequences.** We made the assertion in the previous paragraph that the derived sequence of a cutting sequence is almost constant. While this makes intuitive sense, showing that it must always be the case is an important step towards understanding how the process of finding values of sequences and their derivations can help to determine the slope of the initial line that generated the first cutting sequence. This is the fundamental goal of Theorem 1.8.

The first question we encounter when considering the concept of derived sequences being almost constant is the question "what are we really doing to the sequence when we are deriving it?" The obvious answer is that we are shortening it, but the slightly more sophisticated answer is that we are transforming the square grid. In fact, we are performing a linear transformation on it, taking $ab^{v_0} \to a'$ and $b \to b'$. The 2×2 matrix of the transformation is merely:

$$T = \begin{bmatrix} 1 & 0 \\ v_0 & 1 \end{bmatrix}$$

where again v_0 is the value of the sequence. To verify that this is indeed correct, we can check how this matrix acts on the b and a matrices:

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} and \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

.

We see that:

$$\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

.

This means that the horizontal segment of the grid is altered while the vertical segment stays as it is. We will demonstrate this fact in the example below without using matrix multiplication.

Consider an almost constant sequence of value 2, which would have the following transformation matrix for its derivation:

$$T = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$$

.

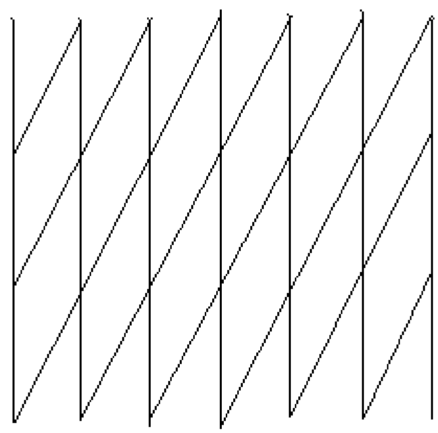
The sequence would look something like:

$$ab^2ab^3ab^2ab^2ab^3\ldots$$

As we can see, this transformation derives according to the rules $ab^2 \to a'$ and $b \to b'$.

If we think about this geometrically, the new grid will consist of a number of parallelograms, as shown in the figure below

FIGURE 5



It is crucial to note that the original line that we were considering remains unchanged because we only changed the grid. This makes the answer to our question simple: the derived sequence of an almost constant sequence is almost constant because it is a cutting sequence through the transformed grid, and every cutting sequence is almost constant as a consequence of Definition 1.4.

1.6. **Slope of Derived Sequences.** When we consider the slope of a derived sequence, we need to first recognize that the slope is relative to the new grid. Instead of calculating "how far the line moves vertically for every shift of 1 horizontally," we need to calculate how far the line moves vertically along the path containing a' segments for each movement of 1 horizontally along b' segments. This explains why the slope of the derived sequence differs from the slope of the original sequence, despite the fact that it appears unchanged by our figure.

Let us consider the same sequence of value 2 as above:

$$ab^2ab^3ab^2ab^2ab^3\ldots$$

Since we determined that the grid for the derived sequence of this series would be as shown in the Figure 5, it becomes trivial to determine the slope. Note that the only change to the grid is that the horizontal lines became lines of slope two. This means that every time the line used to move vertically by λ before , it now moves vertically by $\lambda - 2$, because the grid "follows the line" to some extent and the first vertical distance of 2 traveled by the line serves only to cancel out the change in the grid. More generally, the slope of the derived sequence will become $\lambda' = \lambda - v_0$, where v_0 is the value of the sequence.

It is important to recognize at this point that while we do not know the exact slope of the line (this is what we are setting out to discover), we do have a general sense of it. In this case, the line tends to move vertically a little over twice as far as it moves horizontally. Thus we know that $2 < \lambda < 3$ (more generally $v_0 < \lambda < v_0+1$). This means that our new slope, $\lambda - 2$ is between 0 and 1. What will this mean for the cutting sequence of the derived sequence? If we think back to the very beginning of our discussion, we'll see that we showed that when $\lambda > 1$, the sequence featured more b's than a's, and when $\lambda < 1$, the sequence featured more a's than b's. Thus our new sequence will have more a'*'s* than b'*'s.*

But then what will this mean about it's value? Recall Definition 1.4, which states "any almost constant sequence has a value $[\lambda]$ if $\lambda > 1$, or $[1/\lambda]$ if $\lambda < 1$."

Since in the case of the derived sequence, $\lambda' < 1$, our value is going to end up being $[1/\lambda']$.

1.7. **Putting the Pieces Together - Proving Theorem 1.8 and Explaining the Continued Fraction.**

Proof. Since we now understand all the basic components of how the slope of an almost constant sequence and its derivation can be discovered, it remains only to show that if we put these pieces together, we will get the continued fraction mentioned above. We will assume $\lambda > 1$. This assumption is unnecessary (and we could get around it by flipping a and b at the beginning if we so desired), but is simplifying and loses no generality.

Let us begin with an almost constant and characteristic sequence S of value v_0. By our analysis above, $v_0 < \lambda < v_0 + 1$.

We will now derive the sequence for the first time. Again, as shown in the previous paragraph, the slope of our derived sequence will be $\lambda - v_0 < 1$. Since this number is less than one, we are now going to swap a' and b', obtaining a new line of slope

$$\lambda' = \frac{1}{\lambda - v_0} > 1.$$

Since we have now come full circle (having obtained the slope of the derived sequence in terms of the original slope, and flipped it to make it larger than 1), we now know that we could continue this process indefinitely, and that the next step would give us the slope of the second derived sequence, again flipped to make the slope greater than 1. It would be:

$$\lambda'' = \frac{1}{\lambda' - v_1} = \frac{1}{\dfrac{1}{\lambda - v_0} - v_1} > 1.$$

Let us assume that we continue this process indefinitely, and ultimately find the slope of some very distant derivation of our original sequence. In order to simplify the mathematical equations, we will call S^n the slope of the n^{th} derivation of the original equation. Thus, for example, for n = 5,

$$S^5 = \cfrac{1}{\cfrac{1}{\cfrac{1}{\cfrac{1}{\cfrac{1}{\cfrac{1}{\lambda - v_0} - v_1} - v_2} - v_3} - v_4} - v_5}$$

In order to determine our original λ, we must solve this equation for lambda. By some simple multiplication and addition, we obtain:

$$\lambda = v_0 + \cfrac{1}{v_1 + \cfrac{1}{v_2 + \cfrac{1}{v_3 + \cfrac{1}{v_4 + \cfrac{1}{v_5 + \cfrac{1}{S^5}}}}}}$$

Of course, if we continue this process indefinitely, instead of stopping at the fifth derivation, we'd obtain:

$$\lambda = v_0 + \cfrac{1}{v_1 + \cfrac{1}{v_2 + \cfrac{1}{v_3 + \cfrac{1}{v_4 + \cfrac{1}{v_5 + \dots}}}}}$$

Since the numbers become less relevant to λ as we move farther down the continued fraction, the ultimate slope of the "final" derived sequence is of little importance. Also note that since this is a standard continued fraction, we can write:

$$\lambda = [v_0, v_1, v_2, v_3, v_4 \dots]$$

It is interesting to note that we could also prove that every finite characteristic sequence is linear, though why this fact is true is much easier to show heuristically, as we will do in the following example.

□

1.8. **One Final Detailed Example.** To conclude our section on cutting sequences in euclidean space, we will consider one final example of a finite sequence and solve for its slope. This exercise will be more for the sake of seeing the steps carried out without generality than to show anything new.

Example 1.9. Our sequence will be as follows:

$$S = ab^2ab^3ab^3ab^3ab^2ab^3ab^3ab^2ab^3ab^3ab^3ab^2ab^3ab^3ab^3$$

$value(S) = 2$. $a' = ab^2$, $b' = b$

$S' = a'a'b'a'b'a'b'a'a'b'a'b'a'a'b'a'b'a'b'a'a'b'a'b'a'b' = a'^2b'a'b'a'b'a'^2b'a'b'a'^2b'a'b'a'b'a'^2b'a'b'a'b'$

Now we flip a''s and b''s, calculate a value, and determine a new transformation. for the sake of visibility, let $d = a'$ and $c = b'$, where these refer to the new a' and b' (after the switch)

$$S' = c^2dcdcdc^2dcdc^2dcdcdc^2dcdcd$$

$value(S') = 1$. $c' = c$, $d' = dc = cd$

$$S'' = c'd'd'd'c'd'd'c'd'd'd'c'd'd'd' = c'd'^3c'd'^2c'd'^3c'd'^3$$

Again, we will flip c''s and d''s, calculate a new value, and determine a new translation. Since we will continue this process a few more times, we will no longer write out an explanation like this for each step. Let $f = c'$ and $e = d'$, where these refer to c' and d' after we switch them.

$$S'' = ef^3ef^2ef^3ef^3$$

$value(S'') = 2$. $e' = ef^2$, $f' = f$

$$S''' = e'f'e'e'f'e'f' = e'f'e'^2f'e'f'$$

$value(S''') = 1$
$value(S'''') = 1$
$value(S''''') = 1$
$value(S'''''') = 1$

Thus we obtain:

$$\lambda = [2, 1, 2, 1, 1, 1, 1] = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1}}}}}} = \frac{49}{18}$$

2. Cutting Sequences in the Upper Half Plane

Now that we have shown the equivalence of cutting sequences and the slopes of lines in euclidean space, we will now examine how the concept of cutting sequences can be adapted to they hyperbolic upper half plane, defined as:

$$\mathcal{H} = \{z \in \mathbb{C} | Im(z) > 1\}$$

Recall that in euclidean space, our first step towards obtaining cutting sequences was creating a square grid and labeling its parts. While it would be tempting to follow a similar method in hyperbolic space, this approach makes the problem unnecessarily difficult, because drawing a square grid of equal length sides in hyperbolic space is much more involved (though it is possible). Instead of approaching the problem in this manner, we are going to use the important fact that it does not matter what grid we use, so long as it is well defined and we are able to prove that a given cutting sequence can be transformed to a unique geodesic.

2.1. **Introducing $SL(2,\mathbb{Z})$.** $SL(2,\mathbb{Z})$, also known as the modular group, is defined as follows:

Definition 2.1. $SL(2,\mathbb{Z}) = \{(a,b,c,d) | a,b,c,d \in \mathbb{Z}, ad - bc = 1\}$

Elements of $SL(2,\mathbb{Z})$ are used to generate isometries in hyperbolic space using Mobius transformations. In other words, if we take an element from $SL(2,\mathbb{Z})$ consisting of (a,b,c,d), and insert it into the equation

$$z \to \frac{az+b}{cz+d}$$

we obtain an isometry, or distance-preserving transformation, in hyperbolic space. Note that composition of isometries corresponds directly to multiplication of matrices.

In order to create our grid, we are going to use a tesselation of the fundamental region of $SL(2,\mathbb{Z})$ that is associated to $\Gamma(2)$, a subgroup of index 6 in $SL(2,\mathbb{Z})$. We will associate sides of the fundamental region according to the maps:

$$A : z \to \frac{-1}{z}$$
$$B : z \to 2 - \frac{1}{z}.$$

Finally, these regions can be rotated by $2\pi/3$ about the point $1 + \frac{\sqrt{3}i}{2}$ onto one another by the matrix

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

It is important to note that these specifics are, as previously stated, not relevant to the basics of cutting sequences, but are provided as a means to explain the grid of triangles depicted in Figure 6.

2.2. **Considering our Triangle Grid.** When we look at Figure 6, we see that it consists entirely of triangles with identified sides. If we label the three sides of a given triangle as f, g, and h, we will see that, by the rotation V, we will be able to label the sides of every one of these triangles in much the same way that we labeled sides in euclidean space.

Now that we have tesselated hyperbolic space and labeled all the lines that we used to create our "grid", we could begin to create cutting sequences. We would begin by drawing a geodesic from any point in the space down to a point on the horizontal axis. Keep in mind that since we are in hyperbolic space, geodesics look

Figure 6

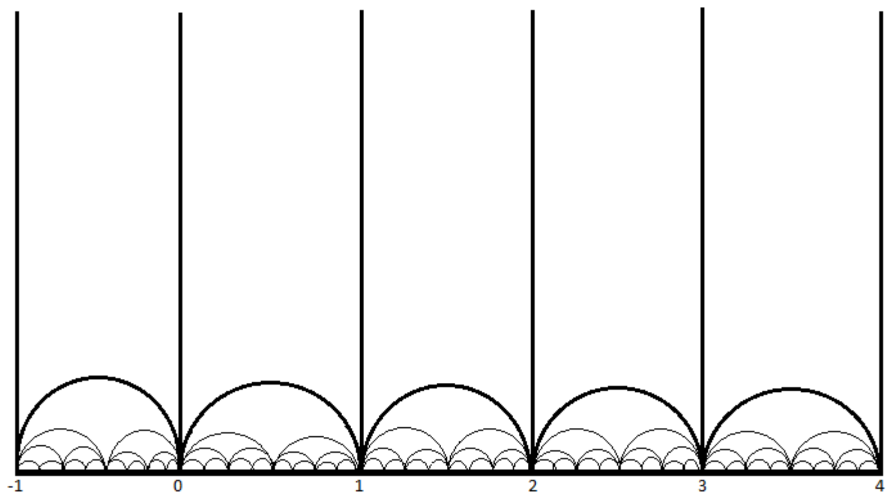
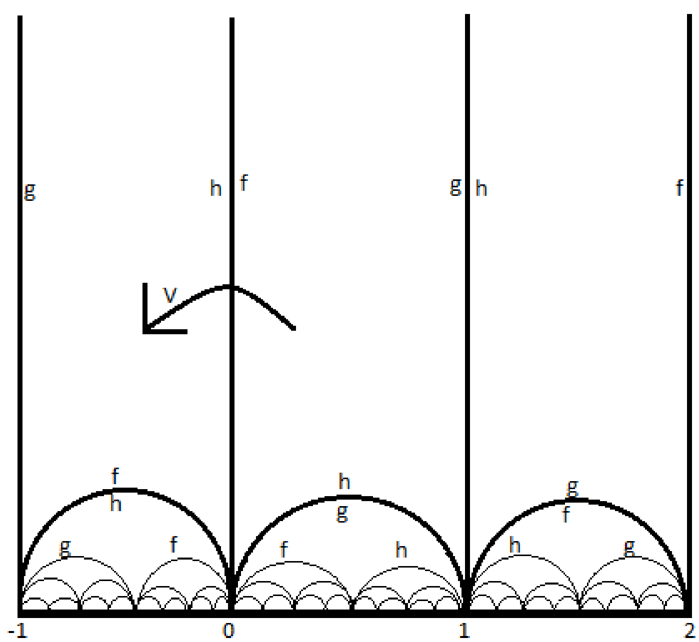


Figure 7



like the arcs of circles centered on the horizontal axis. We would then catalogue which combination of letters, f's, g's, and h's show up in our patten and obtain a cutting sequence in the same way we obtained euclidean cutting sequences.

The first, obvious thing to notice is that this cutting sequence would be "bad" because it would consist of three different letters. This is needlessly complicated, since we did not need three different letters in euclidean space, so it follows that we probably do not need three letters in hyperbolic space. Thus our first step will be to simplify this to a cutting sequence of only two letters.

In order to do this, we will begin by considering what happens to a triangle when a geodesic passes through it. If the geodesic passes through two sides of the triangle, it will cut the triangle into another triangle and a quadrilateral. Based on the way our figure is drawn, it is extremely difficult to construct a geodesic that crosses through a vertex (this only happens at infinity in either direction), but if this occurs, we create two triangles.

We can then create an alternate formulation of the cutting sequence by labeling the path of the geodesic by "what it leaves in its wake." If a geodesic begins by passing through one triangle, we will label it either L or R based on whether it leaves a triangle on its left or right. We can continue to do this as it passes through more and more triangles until it hits the axis. This will allow us to obtain a new cutting sequence that consists only of L and R's. Note that this new sequence can be directly related back to our original sequence of f, g, and h elements by a simple conversion of pairs of letters into the letter L or R that they convert to. This is possible because every triangle has the same sides in the same places, so a line crossing through two given sides will always split the triangle the same way, with the triangle always being on one side and the quadrilateral always on the other. The relationships between f, g, h and L, R cutting sequences are as follows:

$$fh \to R$$
$$gf \to R$$
$$hg \to R$$
$$hf \to L$$
$$fg \to L$$
$$gh \to L$$

2.3. **Geodesic Equivalence.** We will now make our first large claim:

Claim 2.2. *Two geodesics in $\mathbb{H}$ are equivalent in $SL(2,(Z))$ (which means that one can be transformed into the other by an isometry in $SL(2,\mathbb{Z})$) iff their L, R cutting sequences are the same.*

While this may initially seem absurd, we come to realize how trivial this is when we think about how our grid is actually infinite, and not finite as we have drawn it. As we get closer and closer to the horizontal axis, more and more triangles that appear tiny, but in fact are not, spring up. Just like a line in the Euclidean grid, a geodesic in Hyperbolic space will go through an infinite number of these triangles.

Once we realize this fact, the idea that a cutting sequence means that two geodesics are the same seems almost obvious. For a geodesic to have a given cutting sequence, it has to cut through these triangles in all the right places. Our triangles close to the horizontal axis, which appear small, will make this very difficult, since the geodesic will, cut a very specific sequence, which pins its path down to a smaller and smaller set of possibilities. As this specification on where it cuts through triangles increases with the new elements of our cutting sequence, it is like we are decreasing the potential for variance of the line from its trajectory. Since we have an infinitely long cutting sequence, we are, in a sense, taking the limit of this physical line as it approaches the horizontal axis (which is the same as "as we consider more and more elements of its cutting sequence"), and this limit is defined and unique. This means that if we know that a given cutting sequence is the cutting sequence of a geodesic, and we know that another geodesic shares this cutting sequence, they are both the same, unique geodesic, though perhaps transformed by some V in $SL(2, \mathbb{Z})$.

2.4. **Reintroducing Continued Fractions.** While the previous result is certainly interesting, it is definitely not the most we can do with our triangle grid. We will conclude with a proof of the following theorem:

Theorem 2.3. *Let μ be a positive real number, and let $\gamma(\mu)$ be a geodesic ray beginning at a point on the imaginary axis and traveling to μ. We obtain the L, R cutting sequence of this ray, which will look like*

$$L^{v_0} R^{v_1} L^{v_2} \dots$$

for $\mu > 1$ and like

$$R^{v_1} L^{v_2} R^{v_3} \dots$$

for $\mu < 1$. Then $\mu = [v_0, v_1, v_2 \dots]$, which is the standard continued fraction. If there is no v_0, as in the second case above, $v_0 = 0$.

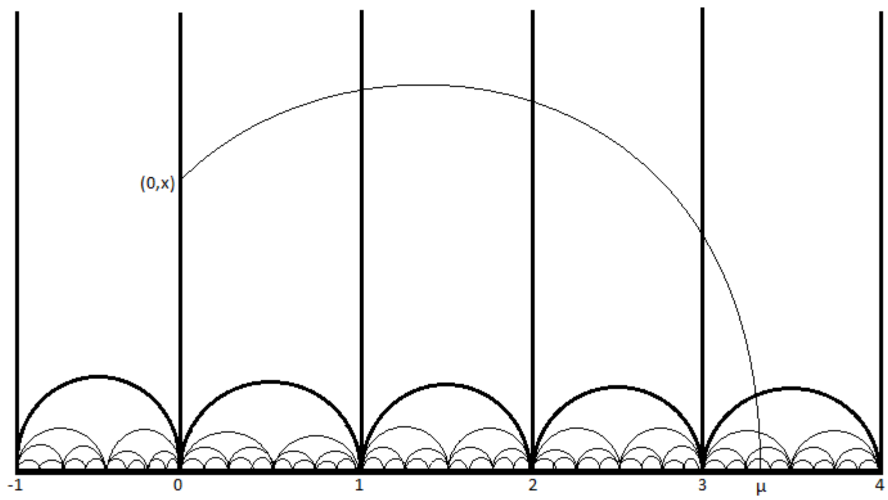
This result is, to put it simply, quite surprising. It says that if we have a point and we know a geodesic ray traveling away from that point and it's cutting sequence, we can know exactly where the point will end up on the horizontal axis. It's interesting to note that if we know only some of the cutting sequence of the geodesic, which is more likely in practical applications, we will still be able to predict a region in which the endpoint of the geodesic will fall. The more information we know about the cutting sequence, the more precise this region will become.

It is also interesting to note that, as in euclidean space, this theorem allows us to redefine what it is we "need" to describe a unique line. Normally in hyperbolic space we need three points to define a unique line, but with this theorem, we show that we don't need two of those points, but instead only need one point on the plane and the cutting sequence of the geodesic ray traveling away from that point in either direction.

Proof. Let us begin by considering a geodesic ray starting at the coordinate $(0, x)$, where $x > 0$ because the point is in the Upper Half Plane, and traveling to the right towards the axis. We will say that it hits the horizontal axis at the point μ, as shown in the figure below.

The first thing to notice with this ray is that its path could follow one of two possibilities. Either it will hit the axis before 1, or it will hit the axis after 1. Note that if it hits the axis before 1, then the first element of the cutting sequence will be an R. Similarly, if it hits the axis after 1, it's first element will be an L. In fact, it will have a number of L's at the beginning of it's cutting sequence corresponding directly to how far it travels horizontally before falling down towards the axis. For example, in Figure 8 the ray ends between 3 and 4, and the cutting sequence of the ray consists of 3 $L's$ followed by an R. Hence our observation:

FIGURE 8



Observation 2.4. Let γ be a geodesic ray beginning at (0,x) in the upper half plane and traveling to the right. Let v_0 be the number of $L's$ at the beginning of γ's L, R cutting sequence. Then the point μ at which γ hits the horizontal axis has the property:

$$v_0 < \mu < v_0 + 1$$

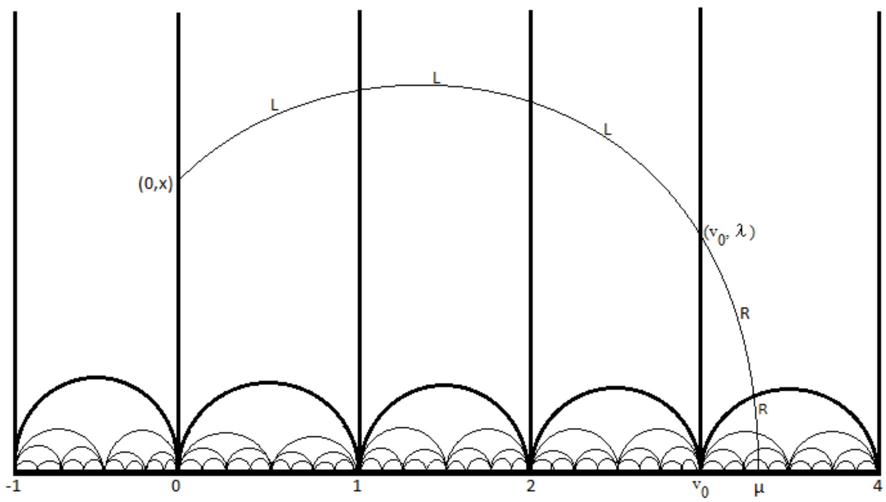
This observation is quite trivial when we consider what it says and what these geodesics look like on the upper half plane, but it is a very important first step in our proof. After we have completed this step and learned that $v_0 < \mu < v_0 + 1$, we will define a new variable λ, which is the value that makes (v_0, λ) a point on ray γ.

Now it appears that the next step would be a similar attempt to find some trick to give us more information about the line by using more of its cutting sequence, and indeed this is the case. However, if we attempt to consider the ray now, as it plummets towards the horizontal axis from the point (v_0, λ), it is difficult to discern any obvious pattern in its cutting sequence. For this reason, we will first transform the rest of our ray according to the map:

$$\eta_1 : z \to -\frac{1}{z - v_0}$$

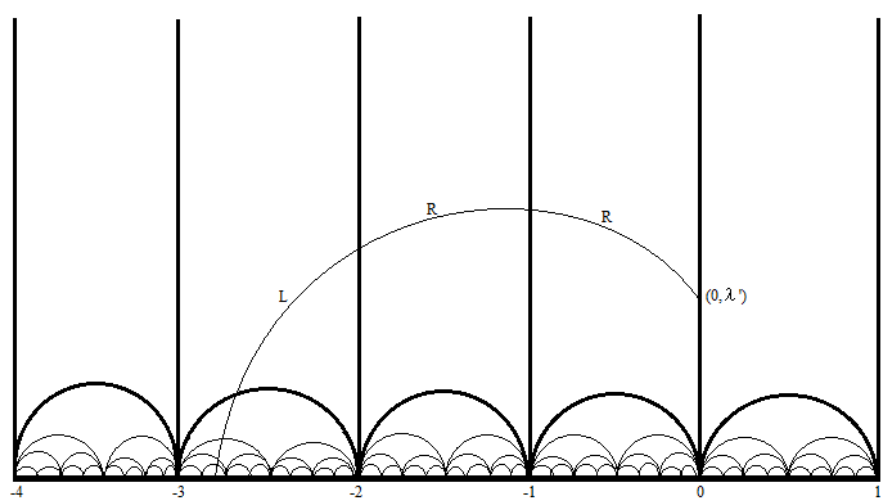
Note the similarity of this map to our original map A in the definition of our space. This is because we are using this transformation to both isometrically rotate the space according to the map A while also shifting it back to the vertical axis. More specifically, this mapping will map our point (v_0, λ) to a point on the imaginary axis, say $(0, \lambda')$ and will shift the geodesic ray to a ray pointing to the left and ending at the point $-\frac{1}{\mu - v_0}$. Now that we have made this mapping, the next step becomes exactly the same as our previous step, except with the letter R instead of

FIGURE 9



L, because the new geodesic is traveling to the left. This step is depicted in Figure 10.

FIGURE 10



Note that in this figure, there are two R's before an L, so $v_1 = 2$.

Now, if we were to repeat this process, we would obtain another isometric transformation of the original geodesic that would again start at the imaginary axis and travel to the right. Counting the number of $L's$ at the beginning of this new sequence would allow us to obtain v_2. It becomes a relatively simple exercise to continue this process indefinitely to obtain the infinite string of values, but this would require continually redrawing the figure and performing the transformation of the geodesic each time. Instead, we will take a step back and consider what exactly we are doing to v_0 to find v_1, to v_1 to v_2, and so on.

As stated before, our first transformation is

$$\eta_1 : z \to -\frac{1}{z - v_0}$$

Note that the endpoint of this new geodesic that we drew was at the point $-\frac{1}{\mu - v_0}$, and that this is equal to $v_1 + k_1$, where $v_1 = [\frac{1}{\mu - v_0}]$ and k_1 is the decimal part of $[\frac{1}{\mu - v_0}]$. This means that:

$$v_1 + k_1 = \frac{1}{\mu - v_0}$$

Where $0 < k_1 < 1$

Solving for μ, we obtain:

$$\mu = \frac{1}{v_1 + k_1} + v_0 \tag{2.5}$$

Now our next transformation, η_2, would be:

$$\eta_2 : z \to -\frac{1}{z - v_1}$$

This will again lead to

$$v_2 + k_2 = \frac{1}{\frac{1}{\mu - v_0} - v_1} \tag{2.6}$$

If we continue this process multiple times (say 5, for the sake of space), we will obtain:

$$v_5 + k_5 = \frac{1}{\frac{1}{\frac{1}{\frac{1}{\frac{1}{\mu - v_0} - v_1} - v_2} - v_3} - v_4}$$

Now, just like in the euclidean case, we must solve for μ, since our goal was to find the point of intersection of the original geodesic with the horizontal axis. By the same simple multiplication and addition, we obtain:

$$\mu = v_0 + \cfrac{1}{v_1 + \cfrac{1}{v_2 + \cfrac{1}{v_3 + \cfrac{1}{v_4 + \cfrac{1}{v_5 + k_5}}}}}$$

As we continue this process indefinitely (traveling further down the cutting sequence, we obtain:

$$\mu = v_0 + \cfrac{1}{v_1 + \cfrac{1}{v_2 + \cfrac{1}{v_3 + \cfrac{1}{v_4 + \cfrac{1}{v_5 + \dots}}}}} = [v_0, v_1, v_2, v_3, v_4, v_5 \dots]$$

Which is our desired result.

□

It is interesting to note that while the math turned out to be very similar to the math for euclidean space, our method was quite different. Instead of deriving the sequence at each step for the entire sequence, we instead worked our way down the sequence, starting with "how many L's were at the beginning", then "how many R's are at the beginning of what remains after we move past the L's", and so on. However, both of these methods derive directly from our use of cutting sequences, and are merely the best way for us to simplify each individual problem using an infinite series of simple and repeatable steps.

While we have made a significant amount of progess in our study, this is only the beginning of what can be achieved with cutting sequences. When we begin to consider cutting sequences on modular surfaces, we see that they have practical applications as well, allowing us to determine information about geodesics on higher dimensional objects that we otherwise have no simple way to obtain, using only the knowledge of the isometries between the these objects and simpler geometries like the two-dimensional upper half plane.

Acknowledgments. It is a pleasure to thank my mentor, Alex Wright, for helping me edit and finalize the paper, while also providing a fresh and knowledgable perspective on the proofs and examples.

References

[1] Caroline Series, The Geometry of Markoff Numbers, The Mathematical Intelligencier Vol 7, No 3. 1985.

[2] Svetlana Katok, Fuchsian Groups, University of Chicago Press. 1992.