The Geometry of Markoff Numbers

Caroline Series

Markoff Irrationalities

It is well known that any irrational number θ can be approximated by a sequence of rationals p_n/q_n which are "good approximations" in the sense that there exists a constant c so that $|\theta - p_n/q_n| < c/q_n^2$. The rationals p_n/q_n are of course the *convergents*, or nth step truncations, of the continued fraction expansion

$$n_0 + \cfrac{1}{n_1 + 1/n_2 + \cdots} = [n_0, n_1, n_2, \ldots] \text{ of } \theta.$$

It is natural to ask for the least possible value of c, in other words, for given θ, find

$$\nu(\theta) = \text{Inf}\{c: |\theta - p/q| < c/q^2 \text{ for infinitely many } q\}.$$

It turns out that $\nu(\theta) \leq 1/\sqrt{5}$ with equality only if θ is a "noble number"† whose continued fraction expansion ends in a string of ones. In 1879 Markoff improved this result by showing that there is a discrete set of values ν_i decreasing to $1/3$ so that if $\nu(\theta) > 1/3$ then $\nu(\theta) = \nu_i$ for some i [8].

The numbers ν_i are called the *Markoff spectrum* and the corresponding θ's, *Markoff irrationalities*. Markoff irrationalities have continued fraction expansions whose tails satisfy a very special set of rules, often called the Dickson rules [4]. The tail 1,1,1, . . . is the simplest example. What these rules are will become clear as we proceed. Markoff gave a prescription for determining all of these irrationalities starting from the solutions of a certain diophantine equation and linked his results to the minima of associated binary quadratic forms.

Recently there has been a revival of interest in this topic, starting from the realisation that each ν_i together with its corresponding class of Markoff irrationalities is associated to a simple (non-self-intersecting) loop on the punctured torus, as shown in Figure 1.

The details have been worked out most fully by A. Haas [6], based on earlier work of Cohn [2, 3], and Schmidt [10]. Lehner, Scheingorn and Beardon [7] tackle the same problem but base their analysis on a sphere with four punctures.

It turns out that almost all the results follow from some rather simple observations about the way in which straight lines cut certain tessellations of the Euclidean and hyperbolic planes and it is these ideas which we want to explain here. Before understanding approximations we shall need to make a fairly lengthy digression to investigate such cutting patterns, for which I offer no apology, for the approach via the patterns is quite as fascinating as Markoff's theory itself.

The Square Grid

Let us begin with a problem in Euclidean geometry. Take the square grid Λ in Figure 2 and label vertical sides by a and horizonal sides by b. Let L be any straight line in the plane, for definiteness directed into the positive quadrant. Walking along L one meets the sides a,b in a certain sequence, *babbabbbabbabbab* in the diagram, which we shall call the *cutting sequence* of L. (If L goes through a vertex, record the sequence as either *ab* or *ba*.) The problem which we pose is this: *describe precisely which sequences of a's and b's occur as cutting sequences of lines in the plane.*

Caroline Series

Figure 3 depicts L drawn horizontally with the positions of the a's and b's marked along its length. If β denotes the distance along L between two vertical segments and α is the distance between horizontal segments then λ = slope $(L) = \beta/\alpha$. If $\lambda > 1$ we make two observations:

Observation 1. The appearances of a are *isolated*, that is, between any two a's is at least one b.

Observation 2. Between any two a's there are either $[\lambda]$ or $[\lambda] + 1$ b's. (Here $[\lambda]$ is the integer part of λ.)

Of course, if $\lambda < 1$, the roles of a and b are reversed, and in observation 2 we read $1/\lambda$ for λ. If L were directed into some other quadrant we would replace a by a^{-1} and b by b^{-1} as appropriate.

Let us call any sequence of a's and b's satisfying 1 and 2, whether or not it is obtained as the cutting sequence of some L, *almost constant*, and call the exponent $[\lambda]$ or $[1/\lambda]$ its *value*.

Given any almost constant sequence s of value n, set $a' = ab^n$, $b' = b$. It is clear that we can rewrite as a sequence s' in the symbols a', b', called the *derived sequence* of s. Of course, s' may itself be almost constant, in which case we may derive it again. Call a sequence which can be derived arbitrarily many times, *characteristic*.

The solution to our problem is now remarkably neat: *The cutting sequence of a line L is characteristic and the values of the successive derived sequences are* $n_0, n_1, n_2, \ldots$ *where* λ = *slope* $(L) = [n_0, n_1, n_2, \ldots]$.

Here is an example of such a sequence: $a^3ba^2ba^2ba^2ba^3ba^2ba^2ba^2ba^3ba^2ba^2ba^2ba^3ba^2ba^2ba^2ba^3ba^2ba^2ba^3ba^2ba^2ba^2ba^2b$. It corresponds to a line of slope $[0,2,4,3,2] = 30/67$.

The beautiful patterns obtained in this way seem first to have been noticed by Christoffel [1] and H. J. S. Smith [11]. I was first introduced to them by D. H. Fowler in connection with Greek mathematics, of which more below. The procedure which we have described, based on deriving almost constant sequences, was discovered (or rediscovered?) by E. C. Zeeman [12].

We have already observed that the cutting sequence for a line of slope $[n_0, n_1, \ldots]$ is almost constant of value n_0. Why does the derived sequence have value n_1? The derivation $a' = ab^{n_0}$, $b' = b$ is really a linear map $\Phi = \begin{pmatrix} 1 & n_0 \\ 0 & 1 \end{pmatrix}$ of the plane which takes the square grid Λ to the grid $\Phi(\Lambda)$ of parallelograms in Figure 4. It is not hard to convince oneself by examining Figure 4 that the cutting sequence of L relative to $\Phi(\Lambda)$ is nothing

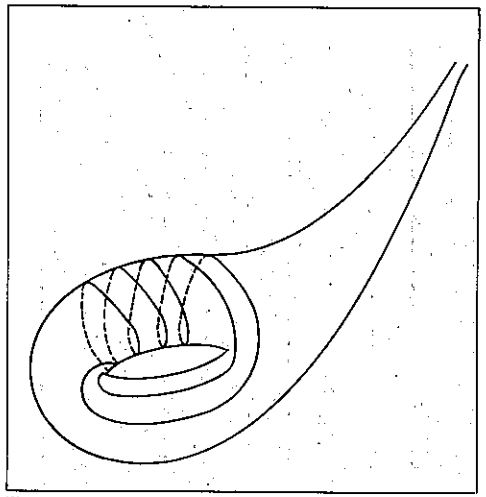

Figure 1. A simple curve on the punctured torus.

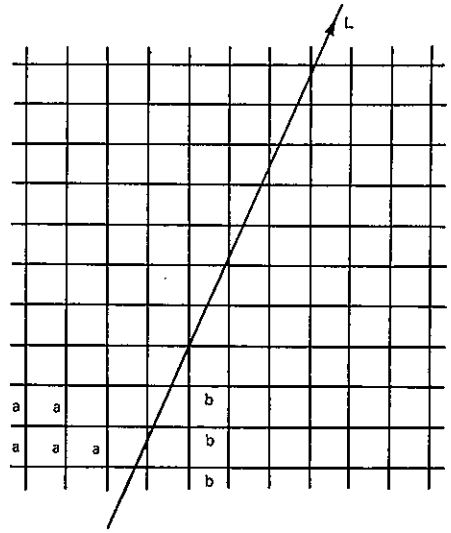

Figure 2.

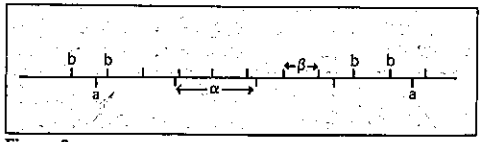

Figure 3.

† This term was invented by I. C. Percival. Noble numbers are those numbers whose tails agree with that of the golden ratio

$$\frac{1+\sqrt{5}}{2} = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdots}}$$

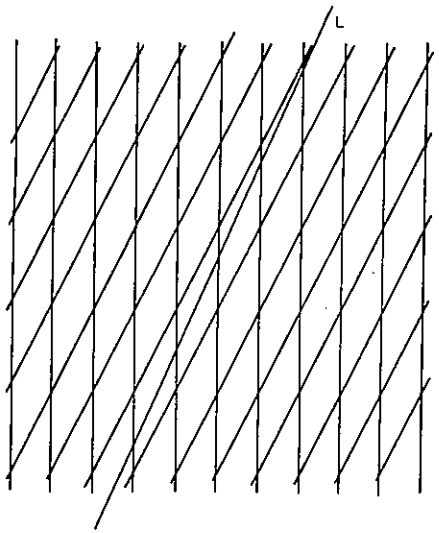

Figure 4.

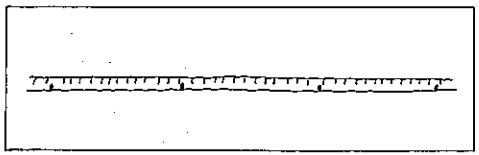

Figure 5. Tally marks showing lunar months and solar years.

other than the derived sequence of the cutting sequence of L relative to Λ. This derived sequence s', being itself a cutting sequence, is also almost constant. We can compute its value by observing that s' is, of course, the same as the cutting sequence of $\Phi^{-1}(L)$ relative to Λ. Now $\Phi^{-1}(L)$ has slope $\lambda - n_0$, and $\lambda - n_0 < 1$. Thus in s' the roles of a and b are interchanged and b' is isolated. Reflecting in the line $x = y$ interchanges a' and b' and gives a line of slope $1/\lambda - n_0 = [n_1, n_2, \ldots]$, from which point the argument repeats.

Does every characteristic sequence occur as the cutting sequence of a line? Not quite; for example, the sequence $b^*a\ b^*$ has an unfortunate "blip" in the middle. However, it is easy to see that every finite characteristic sequence is *linear*, that is, it comes from a line, for the sequence of derivations eventually terminates in a single symbol a^n or b^n which is obviously the cutting sequence of a line relative to some derived grid. Applying in succession the inverse derivations we obtain a line segment with the given sequence as its cutting sequence.

Characteristic sequences are nothing other than the limits of linear ones.

Lunar Cycles

Let us pause for a moment and digress to that most ancient of sciences, astronomy. The patterns of occurrences of one heavenly event relative to another, patterns which must surely have been observed from earliest times, provide natural examples of our cutting sequences. For example, in some years twelve new moons would have been observed, in others thirteen. One could well imagine this data recorded by a sequence of tallies along a rod, perhaps as in Figure 5. What more natural question to ask than what is the pattern of tallies which appear? Of course, the anomalies, or irregularities of the heavens, mean that in fact the interval between two like events is never exactly fixed, so that the tally sequence would deviate slightly from any cutting sequence based on two fixed lengths. A calendar based on the assumption of equal intervals would gradually drift away from observation. Nevertheless, David Fowler has speculated that Plato and Eudoxus might have studied the theoretical properties of tally sequences, and perhaps even the problem of relating tally sequences to continued fractions. This is not so unlikely as it sounds when one recalls that the procedure for expressing a number as a continued fraction is closely related to the Euclidean algorithm. The reciprocal subtraction process used in the algorithm was called by the Greeks *anthiphairesis*, and is thought by David Fowler to be the basis of a pre-Eudoxan theory of proportion [5].

Some ancient calendars in fact embody astonishingly accurate astronomical data. For example, in the calendar called the Metonic cycle, found in Babylonia from around 490 B.C. and introduced to Athens by Meton in 432 B.C., one finds the approximation 19 years = 235 months = 6940 days. This gives a mean synodic month of 29.5319 days, compared to the modern value of 29.5305 days. Incidentally, the number 19 is to be found at the back of the Book of Common Prayer in the formula for calculating the date of Easter, and reaches us via the Jewish calculations for Passover. The ratio 19:235 was used in the gearing of the Antikythera Mechanicism, a remarkable clockwork calendar dating from about 80 B.C. It can in fact be derived from much cruder data than that in the relevant tally sequence and the continued fraction method.

The Punctured Torus

Leaving the Greeks to their anthiphaireses, let us move on some 2,000 years to hyperbolic geometry. Our original problem has, of course, an analogue in the hyperbolic plane. Taking one of the basic squares

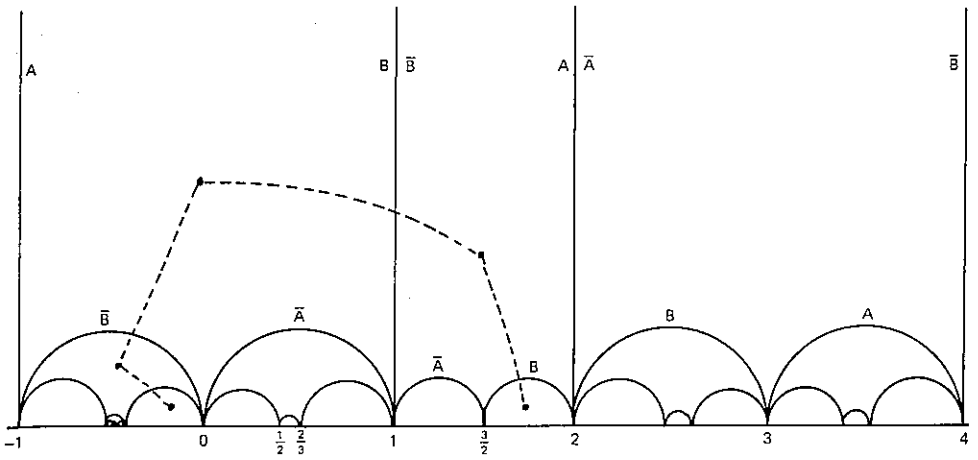

Figure 6. The hyperbolic grid $\mathcal{L}$.

in A and glueing the a and b sides, one obtains a torus. We can think of these glueings as implemented by maps a: $(x,y) \to (x + 1, y)$ and b: $(x,y) \to (x, y + 1)$. (Incidentally, this explains why we chose to label the sides in Figure 2 as we did.) Now take the hyperbolic grid ℓ illustrated in Figure 6 and glue the A and B sides, this time by the maps A: $z \to (z + 1)/(z + 2)$ and B: $z \to (z - 1)/(-z + 2)$.† What you get is again a torus, except that since the corners of the "squares" in ℓ are on the boundary of hyperbolic space, one point is missing on the torus and the effect of the hyperbolic metric is to draw out the region around this puncture into an infinitely long spike or cusp as in Figure 1. Just as the maps a,b of the Euclidean plane generate the abelian group $\mathbb{Z}^2$ which is the fundamental group of the torus, so the maps A,B of the hyperbolic plane generate a free group F which is the fundamental group of the punctured torus T^*. Each "square" in ℓ is an image of the central shaded square S under exactly one element of F, and the labelling of the sides in each square is just a copy of the labels in S.

Recall that straight lines or geodesics in $\mathbb{H}$ are semicircles centered on $\mathbb{R}$, or vertical lines. We can pose the same question as before: *Which A,B sequences occur as the cutting sequences of lines across ℓ?* Of course, our sequences may now contain not only the symbols A,B but also A^{-1}, B^{-1} (henceforth written as $\bar{A}, \bar{B}$), depending on the direction in which we cut sides of ℓ.

Observation 3. In a cutting sequence across ℓ a symbol is never immediately followed by its inverse. A sequence with this property is called reduced. The solution to our problem is this time remarkably simple:

With one exception, every reduced sequence occurs as the cutting sequence of some geodesic in $\mathbb{H}$, terminating sequences corresponding to lines beginning or ending at the puncture.

The exception is the periodic sequence $\ldots A\bar{B}\bar{A}BA\bar{B}\ldots$. This corresponds to a loop encircling the puncture, which is a homotopy class with no geodesic representative.

The idea of the proof is to construct a polygonal path in $\mathbb{H}$ whose cutting sequence is the same as that of a given reduced sequence s. This path will consist of line segments joining one square in ℓ to an adjacent one. Each segment is labelled by the side it cuts. Starting from S, we can construct a path whose cutting sequence coincides with s, shown by dotted lines in Figure 6. The fact that s is reduced simply means that the path never doubles back on itself. It is not hard to prove that such paths always converge to two definite distinct points at infinity with the exception of the bad case $(A\bar{B}A\bar{B})^*$. Joining these points one obtains a geodesic whose cutting sequence is exactly s.

The same method shows that *two geodesics have the same cutting sequence if and only if they can be transformed one into the other by an element in F.* Since transformations in F preserve ℓ and its labelling, sufficiency is obvious. Suppose that two geodesics have the same cutting sequence. By applying a transformation in F to one of them we can suppose that both cut the same side of ℓ at the same point in their cutting sequence. Fixing an initial side fixes the edge paths of both sequences, which therefore coincide. It follows that the two geodesics are the same.

† For more details about hyperbolic geometry and tessellations see the author's earlier *Intelligencer* article "Non-Euclidean Geometry, Continued Fractions and Ergodic Theory" in Vol. 4, No. 1, 1982.

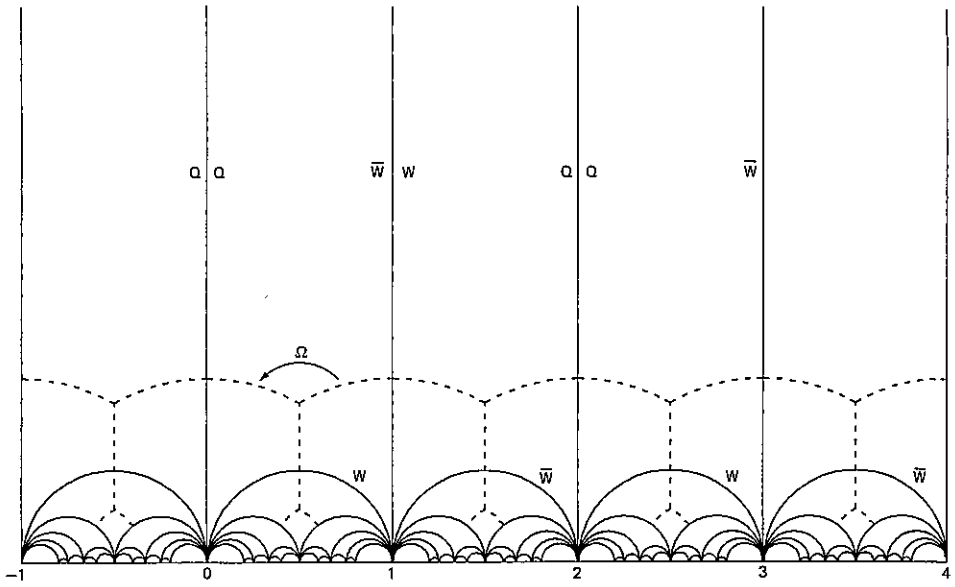

Figure 7. The tessellation Γ subdivided into fundamental regions for SL(2,$\mathbb{Z}$).

SL(2,$\mathbb{Z}$) and Continued Fractions

There was no real reason to take the basic shape S in $\mathfrak{L}$ to be a square. One can play the same game with any tessellation $\mathcal{T}$ provided that the vertices of the fundamental region R all lie at infinity. One such tessellation is illustrated in Figure 7. This is associated to $\Gamma(2)$, a subgroup of index 3 in $SL(2,\mathbb{Z})$.† The sides of the fundamental region R are mapped to each other by the maps Q: $z \to -1/z$, W: $z \to 2 - 1/z$, and this gives the labelling in Figure 7. As shown in the diagram $\mathcal{T}$ is subdivided into three regions each of which is a fundamental region for $SL(2,\mathbb{Z})$. The matrix $\Omega = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ is a rotation by $2\pi/3$ about $1 + \sqrt{3}i/2$ and rotates these regions onto each other.

The cutting sequences of geodesics relative to $\mathcal{T}$ are of the form $\ldots QW^{n_1}QW^{n_2}Q \ldots$, where $n_i \in \mathbb{Z}$. Notice that Q^2 never appears since $Q^{-1} = Q$.

Since $SL(2,\mathbb{Z})$ is generated by Q,W and Ω, its action preserves the tessellation $\mathcal{T}$ although not the Q,W labelling. We can, however, label segments of geodesics cutting across the triangles in $\mathcal{T}$ so as to be invariant under $SL(2,\mathbb{Z})$, by labelling a segment L or R according to whether the vertex of the triangle cut off by the segment is to left or right, as we have done in Figure 8. It is easy to write down a recipe for conversion from Q,W to L,R sequences:

$$\begin{array}{cccccc} QW & \overline{W}\,\overline{W} & WQ & Q\overline{W} & WW & \overline{W}Q \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ L & L & L & R & R & R \end{array}.$$

It now follows that *two geodesics in $\mathbb{H}$ are equivalent under $SL(2,\mathbb{Z})$ if and only if their L,R sequences agree.*

But this is not all. The L,R sequences bring us back to continued fractions! Let θ be any positive real number, and as in Figure 8 let $\gamma(\theta)$ be a geodesic ray joining *any* point on the imaginary axis to θ. Reading off the L,R sequence of $\gamma(\theta)$ we obtain a sequence $L^{n_0}R^{n_1}L^{n_2} \ldots$ (if $\theta < 1$ the sequence begins with R not L). Then $[n_0,n_1,n_2 \ldots]$ is the continued fraction expansion of θ!

The proof is not hard. First, it is obvious that $n_0 = [\theta]$. Let D be the point where $\gamma(\theta)$ cuts $\theta = n_0$. Applying the map τ_1: $z \to -1/z - n_0$, D is mapped to a point D' on the imaginary axis and $\gamma(\theta)$ becomes a ray γ' through D' pointing in the negative direction with endpoint at $-1/\theta - n_0$. The n_1 segments of type R in $\gamma(\theta)$ which follow the initial n_0 segments of type L are now apparent as the n_1 vertical strips crossed by $\tau_1(\gamma)$ before it descends to $\tau_1(\theta)$. Thus $n_1 = 1/\theta - n_0$, so that $\theta = n_0 + 1/n_1 + r$, $0 < r < 1$. Now apply τ_1: $z \to -1/(z + n_1)$ to γ' and proceed as before [9].

† Recall $SL(2,\mathbb{Z}) = \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) | a,b,c,d \in \mathbb{Z},\ ad - bc = 1\}$. Of course $SL(2,\mathbb{Z})$ acts on $\mathbb{H}$ by $z \to az + b/cz + d$.

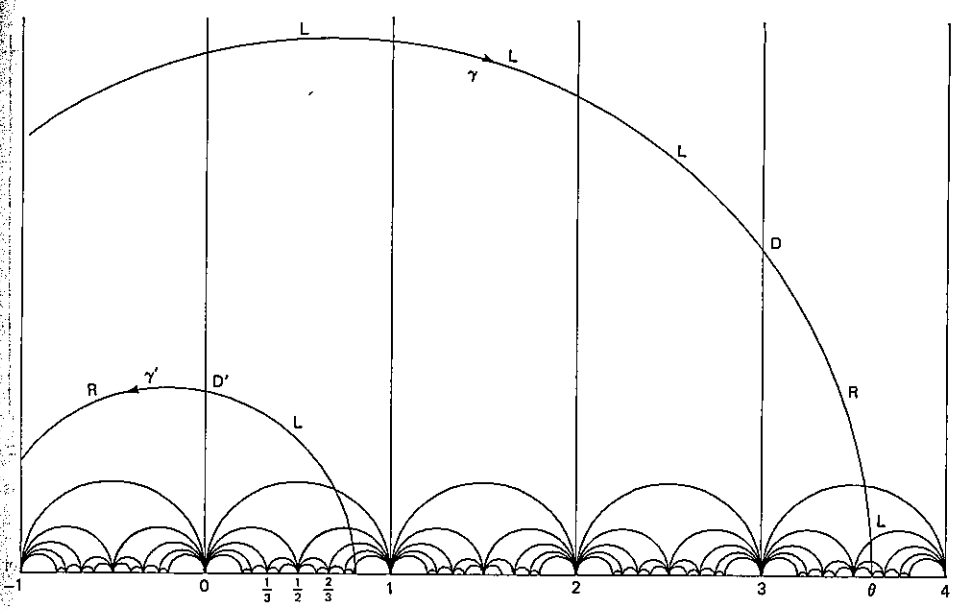

Figure 8. Reading off the continued fraction expansion of θ from $\mathfrak{I}$: $\theta = 3 + \frac{1}{1+\cdots}$.

Simple Curves on the Punctured Torus and the Dickson Rules

We indicated at the beginning that Markoff irrationalities are associated to simple loops on the punctured torus T^*. We are now in a position to understand exactly what these loops are. In fact: *A geodesic on T^* is closed and simple if and only if its cutting sequence is periodic and characteristic.* By the cutting sequence of a curve on T^* we mean, of course, the cutting sequence of any of its lifts to $\mathbb{H}$. Since all these lifts are equivalent under F, we know that cutting sequences coming from different lifts are the same. Closed geodesics correspond exactly to those with periodic cutting sequences. We know that periodic characteristic sequences correspond exactly to lines of rational slope on the square grid Λ. Let L be such a line, and move L if necessary so as to avoid the vertices of Λ.

Since L is disjoint from all its images under the vertical and horizontal translations a and b, its image on T cannot contain any self-intersections; in other words, it is *simple*. Now there is exactly one F-equivalence class of geodesics on the hyperbolic plane $\mathbb{H}$ with the same cutting sequence as L, and it is not hard to show that the corresponding geodesic on T^* is also simple. This geodesic is obtained, if you like, by pulling tight the curve L on T relative to the hyperbolic metric on T^*.

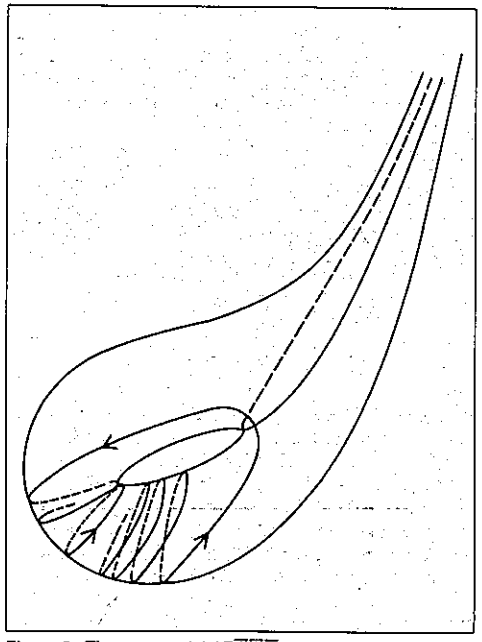

Figure 9. The curve $\cdots AAAB\overline{A}\overline{A}\overline{A}\cdots$.

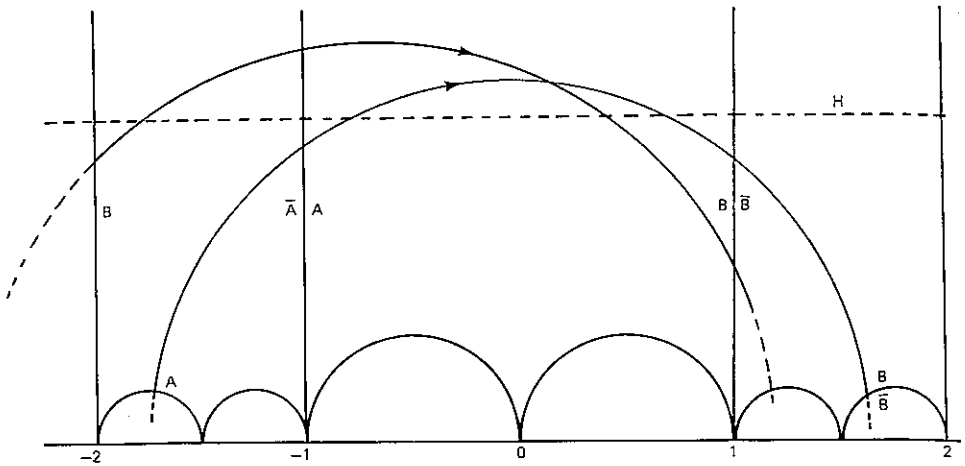

Figure 10. Curves containing $A^2\overline{B}^2$ and $BA\overline{B}$ intersect H.

Now let X be the set of all geodesics on T^* which are limits of simple closed geodesics. Any limit of simple geodesics is simple, and any sequence which is a limit of characteristic sequences is characteristic, for to see that a sequence is not characteristic you only need to look at a sufficiently long finite block. Thus we find that *a geodesic on T^* lies in X if and only if its cutting sequence is characteristic.* Incidentally, there are geodesics on T^* which are simple but do not lie in X. An example is the non-characteristic sequence . . . $AAAB\overline{A}\overline{A}\overline{A}$. . . . This curve spirals into the loop A from opposite directions on the two sides as shown in Figure 9.

Returning to Figures 6 and 7, notice that ℓ is a subtessellation of the $\Gamma(2)$ tesselliation ∇. This means that A,B sequences can also be converted into L,R sequences. The map is as follows:

$A\overline{B}$	AB	AA	$\overline{B}A$	$\overline{B}\,\overline{A}$	$\overline{B}\,\overline{B}$
↓	↓	↓	↓	↓	↓
LL	R	LR	RR	L	RL

and so on. Characteristic sequences have, of course, a very special form: they convert to either . . . $(LR)^nRL\,(LR)^n\overline{RL}$ . . . or . . . $(LR)^nLLRR(LR)^nLLRR$. . . . In addition, these special L,R sequences can be derived in an obvious way and the derived sequences will be of the same kind. Suppose that $\theta \in \mathbb{R}$ is the endpoint of the lift of a simple geodesic α on T^*. The L,R sequences of α and of any ray $\gamma(\theta)$ joining $\mathbb{R}$ to θ will eventually agree. So we can read off the tail of the continued fraction expansion of θ from the L,R sequence of α. It will be $[. . . . 1,1, . . . , 1,2,2,1, . . . , 1,2,2,1, . . .]$, where for some n the blocks of 1's are of length $2n$ or $2(n + 1)$, and all derived sequences are of the same form.

Needless to say, these tails are exactly those which obey the Dickson rules!

Approaching the Puncture

We have established that Markoff irrationalities are exactly the endpoints of lifts of simple geodesics in X. But what have simple geodesics to do with Diophantine approximation? The key is the observation that geodesics in X never approach too close to the puncture P. In fact, in $\mathbb{H}$ they never rise above the line $\mathrm{Im}z = 3/2$. What is even more remarkable is that the converse is also true: *A geodesic γ on $\mathbb{H}$ projects to a geodesic in X if and only if $Imz \leq 3/2$ for all $z \in F\gamma$, where $F\gamma$ is the set of all translates of γ under F.*

For the proof it will be useful to consider two isometries of $\mathbb{H}$ which preserve the lattice ℓ while altering the labelling. One is the reflexion r in the imaginary axis, which interchanges the labels A and B. The other is translation t: $z \to z + 3$, which interchanges A with $\overline{A}$ and B with $\overline{B}$. Notice that γ lies below the line $\mathrm{Im}z = 3/2$ if and only if $\gamma \cap t(\gamma) = \emptyset$.

On the grid Λ the substitution $\tilde{t}$: $a \to \bar{a}$, $b \to \bar{b}$ is implemented by rotation by π about 0. Any line L is either disjoint from or coincides with $\tilde{t}(L)$ and hence L and $\tilde{t}(L)$ project to disjoint or coincident curves on T. But then the same is true of $\gamma(L)$, the corresponding geodesic on T^*. It follows that any curve whose cutting sequence is linear lies below $\mathrm{Im}z = 3/2$. Taking limits, the same holds for the lifts of any curve in X.

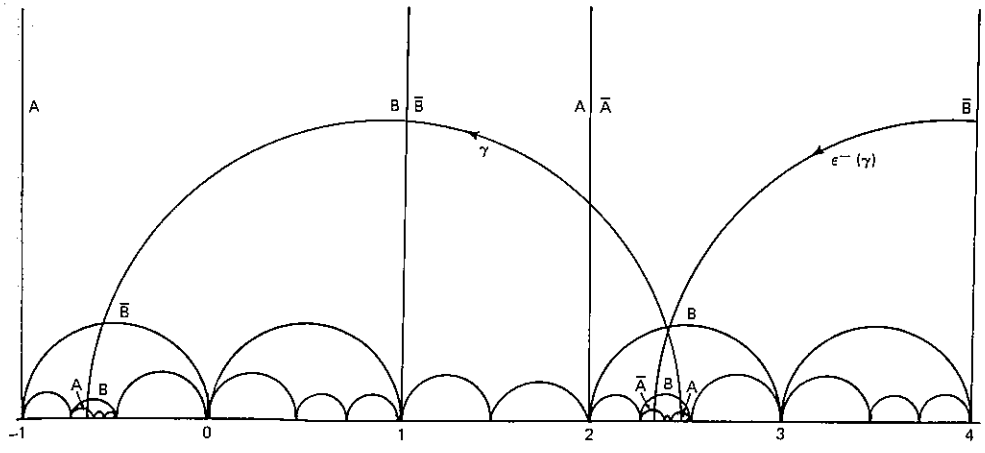

Figure 11. Curves containing $AB^3\overline{A}$ cut H.

The converse statement is the deepest result which we shall prove here, although the methods are really no harder than those we have used to date. First, notice that just as the derivation $a \to ab^n$, $b \to b$ of Λ was induced by a linear map Φ of $\mathbb{R}^2$, so the derivation $A \to AB$, $B \to \overline{A}$ is induced by the isometry δ: $z \to z - 1$ of $\mathbb{H}$. The new set of generators $A' = AB$, $B' = \overline{A}$ of F is associated to a tessellation ℓ' identical with ℓ and with the same pattern of labels but translated one unit to the left.

The cutting sequence of a geodesic γ relative to ℓ' is the derived sequence of the cutting sequences of γ relative to ℓ, where we now extend the meaning of "derived" to mean the sequence obtained substituting $\overline{B}'$ for A and $B'A'$ for B in s.

Suppose that s is a non-characteristic sequence. Using the observations above we may assume that s has been derived enough times that we see in s either a sequence $XY^n\overline{X}$ or X^2YXY . . . XY^2. Let H be the region above $\mathrm{Im}z = 3/2$. Figure 10 illustrates that curves containing sequences $XY\overline{X}$ or X^2Y^2 intersect H.

Suppose inductively that any curve whose sequence contains $XY^n\overline{X}$ has a lift cutting H. Figure 11 illustrates a curve γ containing $AB^{n+1}\overline{A}$ and its image $t(\gamma)$ containing $\overline{A}\,\overline{B}^{n+1}A$. One can see that $\gamma \cap t(\gamma) \neq \emptyset$ unless the sequence $AB^{n+1}\overline{A}$ is preceded by B^nA. But then γ would contain the sequence $\overline{A}B^nA$, which we have already dealt with. Obvious symmetries deal with the other cases.

Using the substitutions r and t we are now reduced to the case where s contains either $A^2BA\,B$ . . . AB^2 or $A^2\overline{B}A\overline{B}$ . . . $A\overline{B}^2$. In the first, the derivation δ produces a sequence containing a block $XY^n\overline{X}$; in the second it gives a sequence all of whose exponents are negative which is, after applying t, already disposed of.

Notice that this proof depends only on looking at finite blocks in the cutting sequence; in other words, to see that a geodesic enters H, it is enough to look at any finite segment along which the derivation rules are violated. This fact will be important to us below.

Diophantine Approximation

We are now finally able to bring all the pieces together and return to Markoff's original approximation problem. What we proved in the last section amounts to showing that a geodesic in X avoids the image $\overline{H}$ of H on T^*. Thus the lifts of such geodesics avoid not only H but all images of H under F. By a simple calculation the image of H under $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in F$ is a circle tangent to $\mathbb{R}$ at a/c, of radius $1/3c^2$. Moreover, as $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ runs through F, a/c takes on all possible rational values. Let us denote the union of all these *horocycles* by N.

We can divide irrational points $\theta \in \mathbb{R}$ into three types, according to the asymptotic behaviour of the L,R cutting sequence of the vertical line $v(\theta)$ joining θ to infinity. Denote the L,R sequence of $v(\theta)$ from the nth term on by $s_n(\theta)$. There are three possibilities:

(i) For some n, $s_n(\theta)$ is characteristic and periodic.
(ii) For some n, $s_n(\theta)$ is characteristic but $s_m(\theta)$ is not periodic for any m.
(iii) $s_n(\theta)$ is never characteristic.

In case (i), $s_n(\theta)$ represents a simple closed geodesic $\overline{\alpha}$ on T^*. Of course $\overline{\alpha}$ lifts to a geodesic α on $\mathbb{H}$ with endpoint at θ. Since the image $\overline{v}(\theta)$ of $v(\theta)$ approaches $\overline{\alpha}$ asymptotically on T^*, and since $\overline{\alpha}$ is a bounded dis-

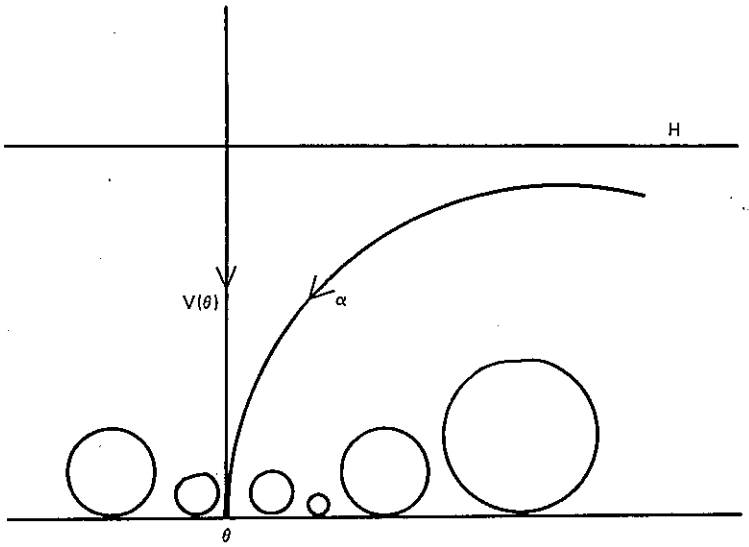

Figure 12. Simple curves avoiding horocycles.

tance outside $\overline{H}$, we see that $v(\theta)$ enters N only a finite number of times. Thus one sees from Figure 12 that the inequality $|\theta - a/c| < 1/3c^2$ has only a finite number of solutions so that $\nu(\theta) > 1/3$.

One can actually calculate that the closest approach of α to $\overline{H}$ is $\log \coth \ell(\alpha)/2$ where $\ell(\alpha)$ is the hyperbolic length of α [6]. This gives an exact value for $\nu(\theta)$.

In case (ii) the tail of $v(\theta)$ is characteristic and hence $\overline{v}$ can be approximated by a sequence of simple curves $\overline{\alpha}_n$. Since there are only finitely many curves with lengths below a given bound, the sequence of lengths tends to infinity and hence the distance to $\overline{H}$ approaches zero. Thus $v(\theta)$ approaches N arbitrarily closely although from some point on it never enters N since $s_n(\theta)$ is eventually characteristic. Combining these facts one sees that $\nu(\theta) = 1/3$.

Finally, in case (iii) the tails $s_n(\theta)$ are never characteristic and so $v(\theta)$ enters N infinitely often. Thus there are infinitely many solutions to $|\theta - a/c| < 1/3c^2$; in other words, $\nu(\theta) \leq 1/3$.

Trace Formulae, Diophantine Equations and Quadratic Forms

Hoping that the reader's patience is not completely exhausted, we will conclude by giving some brief pointers to the connection of our approach to another well known aspect of Markoff's theory, the mimima of binary quadratic forms.

The Markoff spectrum if frequently calculated by introducing *Markoff triples*. These are integer triples (x,y,z) which are solutions of the Diophantine equation

$$x^2 + y^2 + z^2 = 3xyz. \qquad \text{(D)}$$

Associated to such a triple is a pair of real quadratic numbers $\xi,\xi' = 1/2 + y/xz + 1/2(9 - 4/z^2)^{1/2}$. The numbers ξ,ξ' are Markoff irrationalities with $\nu(\xi) = \nu(\xi') = \sqrt{9 - 4/z^2} > 1/3$.

In fact, as explained in [2], Markoff triples are (up to a factor of 3) the traces of triples $(U,V,\overline{V}\,\overline{U})$ such that U,V are a pair of generators for the group F with fundamental region as shown in Figure 13. The simplest solution (1,1,1) corresponds to the A,B generators we used above. The formula (D) is nothing other than one of Fricke's *trace identities* relating traces of matrices in $SL(2,\mathbb{R})$. Starting with the solution (1,1,1) the operations $(x,y,z) \to (z,x,y)$ and $(x,y,z) \to (x,3xy - z,y)$ generate all possible solutions to (D). These operations are the same as the operations of derivation and substitution which we used above.

The geodesics $\overline{\gamma}(A)$, $\overline{\gamma}(B)$ corresponding to the minimal solution (1,1,1) of (D) are simple. Since the operation of derivation is induced by an isometry of T^*, the same is actually true of *all* solutions of (D). Now the geodesic $\overline{\gamma}(M)$ associated to a matrix $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in F$ is the projection of a geodesic $\gamma(M)$ on $\mathbb{H}$ whose endpoints are the fixed points ξ_M,ξ'_M of M on $\mathbb{R}$. These endpoints are of course roots of $c\xi^2 + (d - a)\xi - b = 0$. Thus we see in another way that Markoff irrationalities are the endpoints of lifts of simple geodesics on T^*.

One can associate to M the quadratic form

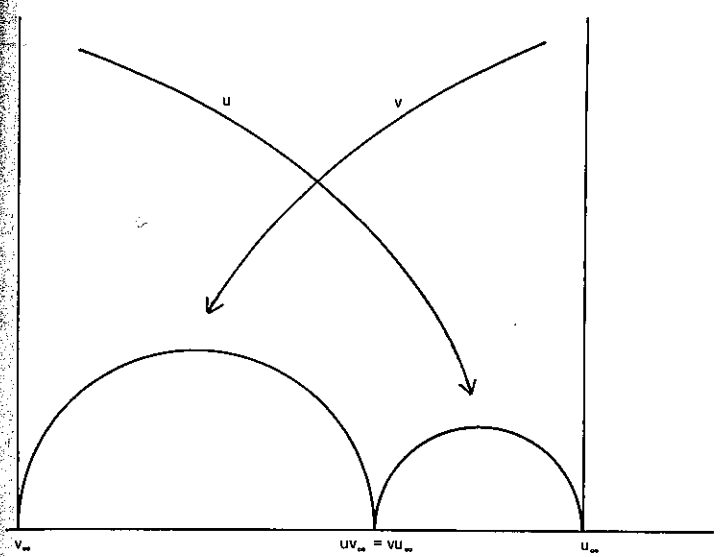

Figure 13. Fundamental region for the punctured torus.

$$\phi_M(x,y) = cx^2 + (d - a)xy - by^2.$$

Since $\gamma(M)$ is simple it lies below the line $\mathrm{Im}\, z = 3/2$, in other words, $|\xi_M - \xi'_M| < 3$. Now $|\xi_M - \xi'_M| = \Delta^{1/2}/Q_M(1,0)$ where $\Delta = \mathrm{Tr}^2 M - 1$ is the discriminant of Q_M, so that $Q_M(1,0)/\Delta^{1/2} > 1/3$.

But we know more. Since the action of $SL(2,\mathbb{Z})$ on $\mathbb{H}$ preserves L,R sequences, it preserves simple geodesics on T^*. Thus if $\gamma(M)$ is simple, so is $\gamma(gMg^{-1})$ for any $g \in SL(2,\mathbb{Z})$. One easily computes that

$$Q_M(x,y) = Q_{gMg^{-1}}(gx,gy)$$

and that Q_M and $Q_{gMg^{-1}}$ have the same discriminant Δ. Given any pair $(x,y) \in \mathbb{Z}^2$ we can always find $g \in SL(2,\mathbb{Z})$ with $(gx,gy) = (1,0)$. Hence

$$Q_M(x,y) = Q_{gMg^{-1}}(1,0) > \Delta^{1/2}/3;$$

in other words,

$$\min_{x,y \in \mathbb{Z}^2} \frac{Q_M(x,y)}{\Delta^{1/2}} > 1/3.$$

Of course the actual value of the minimum can be calculated and is, not surprisingly, $\nu(\xi_M)$. For matrices M which do not correspond to simple geodesics, the minimum lies on or below the value $1/3$.

These are the results of Markoff on minima of binary quadratic forms.

It seems clear from the foregoing that the next level of approximation should be studied by looking at geodesics with one self-intersection. Such geodesics penetrate only a bounded distance into $\overline{H}$. One wonders if these further levels of approximation are perhaps related to phenomena of successive transitions from periodicity into chaos?

References

1. E. B. Christoffel, *Observatio Arithmetica*, *Annali di Mathematica*, 2nd series, *6*(1875), 148–152.
2. H. Cohn, Approach to Markoff's minimal forms through modular functions. *Ann. Math. 61*(1955), 1–12.
3. H. Cohn, Representation of Markoff's binary quadratic forms by geodesics on a perforated torus. *Acta Arithmetica XVIII*(1971), 125–136.
4. L. E. Dickson, Studies in the theory of numbers. Chicago: 1930.
5. D. Fowler, Anthyphairetic ratio and Eudoxan proportion. *Archive for History of Exact Sciences* 24(1981), 69–72.
6. A. Haas, Diophantine approximation on hyperbolic Riemann surfaces, *Bull. A.M.S. 11*(1984), 359–362.
7. J. Lehner, M. Scheingorn, Simple closed geodesics on $H^+/\Gamma(3)$ arise from the Markoff spectrum, preprint.
8. A. A. Markoff, Sur les formes binaires indefinies, I, *Math. Ann. 15*(1879), 281–309; II, *17*(1880), 379–400.
9. C. Series, The modular surface and continued fractions. *J. London Math. Soc.* (1984).
10. A. L. Schmidt, Minimum of quadratic forms with respect to Fuchsian groups I. *J. Reine Angew. Math.* 286/7 (1976), 341–368.
11. H. J. S. Smith, Note on continued fractions. *Messenger of Mathematics*, 2nd series, *6*(1876), 1–14.
12. E. C. Zeeman, An algorithm for Eudoxan and anthiphairetic ratios, preprint.

Department of Mathematics
University of Pennsylvania
Philadelphia, Pennsylvania 19104 U.S.A.