THE GEOMETRY OF MARKOFF FORMS

Andrew Haas
Department of Mathematics
University of Connecticut
Storrs, Connecticut 06268

It is now recognized that the branch of classical number theory pertaining to the study of binary indefinite quadratic forms and their minima enjoys an intimate connection with the theories of Fuchsian groups and hyperbolic Riemann surfaces. This view was pioneered and nurtured by Harvey Cohn through a series of papers he wrote beginning in 1954 [C1-C5], in which the central theme is the relationship between Diophantine approximation and the geometry of the subgroup Γ' of the classical modular group $SL_2(\mathbb{Z})$ and the associated quotient Riemann surface $\mathbb{H}/\Gamma'$.

Asmus Schmidt took a step away from the classical setting of Γ' by redefining the minimum of a quadratic form with respect to an arbitrary zonal Fuchsian group. He succeeded in giving a proof of Markoff's theorem for all zonal Fuchsian groups topologically conjugate to Γ' [Sc]. His proof combines elements of the two classical arguments with Cohns approach via Fricke groups.

The more recent ventures in this spirit have been by Lehner & Sheingorn [L-S], Series [Se] and the author [H]. We share a point of view which emphasizes the role of the space of geodesics on hyperbolic Riemann surfaces.

In this article I will try to motivate and explain the geometric viewpoint.

§1. *Quadratic forms*

Let $\mathcal{F}$ denote the space of binary indefinite quadratic forms; these are functions of two variables

$$f(x,y) = \alpha x^2 + \beta xy + \gamma y^2 \tag{1.1}$$

where α,β , and γ are real numbers, $\alpha^2 + \gamma^2 \neq 0$, and the discriminant $D(f) = \beta^2 - 4\alpha\gamma > 0$. We take as the minimum of a form f the quantity

$$M(f) = \inf \frac{|f(x,y)|}{\sqrt{D(f)}}$$

where the infimum is over all pairs of integers (x,y) excluding $(0,0)$.

Two forms f and g are said to be *equivalent* if either

1. $g = af$ for some nonzero $a \in \mathbb{R}$ or
2. $g = f \circ A = f(ax + by\ ,\ cx + dy)$ for some $\begin{bmatrix} a & b \\ c & d \end{bmatrix} = A \in SL_2(\mathbb{Z})$.

Equivalent forms have the same minima. This is clear for (1), and for (2) it follows if one observes that $A \in SL_2(\mathbb{Z})$ is invertible, preserves the integer lattice $\mathbb{Z}^2$ and preserves the discriminant of a form.

With the above definitions the natural problem becomes one of describing the set of values $\{M(f) \mid f \in \mathcal{F}\}$; and for any value μ to determine those f with $M(f) = \mu$. In light of the equivalence relation on $\mathcal{F}$ this last problem is more manageable if we instead ask for a set of forms with minimum μ so that every form with minimum μ is equivalent to one of those.

The most successful approach to this problem is due to A. Markoff [M]. What he proved was the

Theorem 1.2. *There is a discrete sequence of values M_i , decreasing to 1/3 so that $M(f) > 1/3$ if and only if $M(f) = M_i$ for some positive integer i .*

We refer to a form f with $M(f) > 1/3$ as a *Markoff form.*

Markoff's theorem also provides a means for constructing the values M_i and maximal sets of inequivalent forms with minima M_i from solutions to the Diophantine equation

$$x^2 + y^2 + z^2 = 3xyz \tag{1.3}$$

§2. *Fricke groups*

Let Γ' be the subgroup of $SL_2(\mathbb{Z})$ which is freely generated by the matrices

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} .$$

We call a matrix A a *generator* if there is another matrix B which together with A generates Γ' .

Fricke showed that if A and B generate Γ' then the traces of the matrices A,B , and AB satisfy the equation

$$(2.1) \qquad (trA)^2 + (trB)^2 + (trAB)^2 = (trA)(trB)(trAB)$$

The similarity between (1.3) and (2.1) was observed by Cohn who subsequently developed a new approach to Markoff Theory in terms of the group Γ' . Given a matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ define the form

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

The link with Markoff's theorem is expressed by the

Theorem 2.2. (Cohn [C1]) *f is a Markoff form if and only if f is equivalent to a form f_A where A is a generator of Γ' .*

The spectral values M_i are now constructed in terms of the traces of generators of Γ' .

Schmidt proved a variant of Theorem 2.2 in the context of what he calls extended Fricke groups. He also succeeded in deriving a description of those forms f with $M(f) \geq 1/3$ [Sc] .

A 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL_2(\mathbb{R})$ with real entries and determinant 1 , aside from representing a linear transformation of $\mathbb{R}^2$, acts on the Riemann sphere $\hat{\mathbb{C}}$ according to the rule

$$A(z) = \frac{az + b}{cz + d} .$$

The restriction to real coefficients and determinant 1 guarantees that A is a conformal homeomorphism of $\hat{\mathbb{C}}$ which maps the upper half plane $\mathbb{H}$ onto itself.

When $\mathbb{H}$ is equipped with the metric $ds = \frac{|dz|}{y}$ it becomes the Poincaré model for the (nonEuclidean) hyperbolic plane. The hyperbolic length of an arc $\gamma : [a,b] \to \mathbb{H}$ is then $\ell(\gamma) = \int_a^b \frac{|\dot{\gamma}(t)|}{Im\gamma(t)}\, dt$.

The orientation preserving isometries of $\mathbb{H}$ are precisely the self maps induced by the elements of $SL_2(\mathbb{R})$.

A geodesics in $\mathbb{H}$ is an arc of the form $\gamma = \mathbb{H} \cap C$ where C is a circle or straight line in $\mathbb{C}$ which is orthogonal to the real axis. The easiest way to describe a geodesic in $\mathbb{H}$ is to specify its two endpoints in $\hat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. We let $\mathcal{G}$ denote the space of all geodesics in $\mathbb{H}$.

§3. *Markoff geodesics in hyperbolic space*

To each binary indefinite quadratic form f there is associated a unique hyperbolic geodesic γ_f whose endpoints ξ and ζ are roots of the equation $f(x,1) = 0$ (if $f(x,1)$ is first order, which happens when $\alpha = 0$, then one endpoint is ∞) .

Let ϕ be the map from $\mathcal{F}$ to $\mathcal{G}$ taking a form f to a geodesic γ_f . ϕ is onto, and the preimage of geodesic $\gamma \in \mathcal{G}$ with endpoints ξ and ζ is a one parameter family of forms tf for $t \in \mathbb{R}$ where $f(x,y) = (x-\xi y)(x-\zeta y)$ (if $\zeta = \infty$, $f(x,y) = y(x-\xi y)$) . If we let $\mathcal{PF}$ denote the space of real projective classes of forms in $\mathcal{F}$ then the induced map $\Phi : \mathcal{PF} \to \mathcal{G}$ will in fact be a homeomorphism in the natural topologies on $\mathcal{PF}$ and $\mathcal{G}$.

$SL_2(\mathbb{Z})$ acts on the set $\mathcal{G}$ by treating geodesics as point sets in $\mathbb{H}$ and letting matrices act as Möbius transformations. We may also define an action of $SL_2(\mathbb{Z})$ on $\mathcal{PF}$ by $A(f) = f \circ A^{-1}$. Φ is equivariant with respect to this action.

In order to see this we let Q be a symmetric matrix with $f(x,y) = (x,y)\, Q\, (x,y)^t$. The form $A(f) = f \circ A^{-1}$ is determined by the matrix $(A^{-1})\, Q\, (A^{-1})^t$. It follows that the roots of the equation $f \circ A^{-1}(x,1) = 0$ are $A(\xi)$ and $A(\zeta)$. In particular it follows that two forms f and g are equivalent if and only if there is a transformation $B \in SL_2(\mathbb{Z})$ with $B(\gamma_f) = \gamma_g$.

If A is a matrix in $SL_2(\mathbb{Z})$ with absolute trace greater than 2 then there is a unique geodesic in $\mathbb{H}$ whose endpoints are the fixed points of $A(z)$. When $f = f_A$ the roots of the equation $f(x,1) = 0$ are exactly the fixed points of $A(z)$, and therefore γ_f is the fixed axis of A .

When f is a Markoff form we refer to γ_f as a *Markoff geodesic*.

Define the height of a geodesic γ with endpoints ξ and ζ by

$$ht(\gamma) = \begin{cases} \frac{|\xi-\zeta|}{2} & \text{if } \zeta \text{ and } \xi < \infty \\ \infty & \text{otherwise} \end{cases}$$

The quantity $H(\gamma) = \sup\{ht(g(\gamma)) \mid g \in \Gamma'\}$ is constant on the Γ' orbit of the geodesic γ and is related to the minimum of a form by way of

Lemma 3.1. (Cohn [C2]). $H(\gamma_f) = \frac{1}{2M(f)}$.

In order to clarify the relationship between M and H we will argue the lemma.

Proof: There is no loss of generality if we restrict attention to forms with discriminant 1. If $\alpha \neq 0$ in (1.1) then the endpoints of γ_f are $\frac{-\beta+1}{2\alpha}$ and $\frac{-\beta-1}{2\alpha}$, and $ht(\gamma_f) = \frac{1}{2|\alpha|}$.

Let $\langle(x_i,y_i)\rangle$ be a sequence of points in $\mathbb{Z}^2$ with $\lim_{i\to\infty} |f(x_i,y_i)| = M(f)$. Choose $A_i \in SL_2(\mathbb{Z})$ with $A_i\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}x_i\\y_i\end{bmatrix}$ and set $f\circ A_i = g_i = a_ix^2 + b_ixy + c_iy^2$. Then $\lim_{i\to\infty} |g_i(1,0)| = M(f)$, from which it follows that $M(f) = \lim_{i\to\infty} |a_i| = \lim_{i\to\infty} \frac{1}{2ht(\gamma_i)}$; where $\gamma_i = \gamma_{g_i}$. Therefore, $H(\gamma_f) \geq ht(\gamma_i) = \frac{1}{2|a_i|} \geq \frac{1}{2M(f)}$.

If the inequality were strict then there would exist a form $g = ax^2 + bxy + cy^2$ with discriminant 1 equivalent to f with $ht(\gamma_g) > \frac{1}{2M(f)}$. As we observed above $ht(\gamma_g) = \frac{1}{2|a|}$. It follows that $|g(1,0)| = |a| < M(f) = M(g)$ contrary to the definition of $M(g)$.

If $\alpha = 0$ then we clearly have both $H(\gamma_f) = \infty$ and $M(f) = 0$. □

An immediate consequence of the lemma is a characterization of Markoff geodesics in terms of the height function:

(3.2) *γ is a Markoff geodesic if and only if $H(\gamma) < 3/2$.*

This is the first step in a geometrization of the Markoff theory.

§4. *Markoff geodesics on the punctured torus*

The group Γ' acts discontinuously on the upper half plane and the quotient space $\mathbb{H}/\Gamma' = T_z$ is a hyperbolic Riemann surface which is homeomorphic to a torus from which a point has been removed. This is easily seen by looking at a standard fundamental domain F for the action of Γ' on $\mathbb{H}$. See figure 4.1. The generators $\begin{bmatrix}1 & 1\\ 1 & 2\end{bmatrix}$ and $\begin{bmatrix}2 & 1\\ 1 & 1\end{bmatrix}$ pair opposite sides of F. This is much like the identification of opposite sides on a rectangle to produce a torus except that now the vertices are out at infinity in hyperbolic space; thus the puncture.

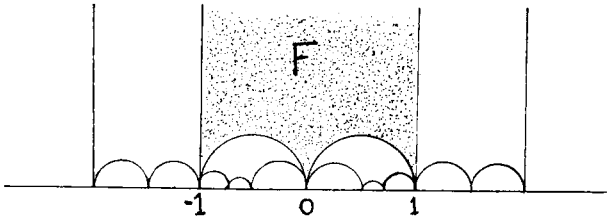
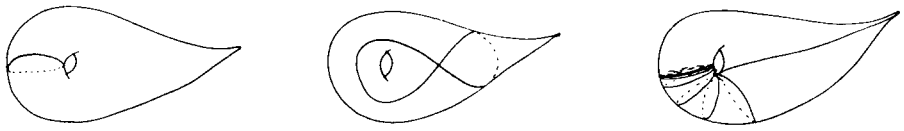


Fig. 4.1. The fundamental domain F for Γ' (shaded) with adjacent Γ'-translates of F.

A geodesic $\tilde{\gamma}$ in $\mathbb{H}$ projects to a geodesic γ on T_z. Every geodesic on T_z arises in this fashion. We say a geodesic γ on T_z is *closed* if it is the projection of a geodesic $\tilde{\gamma}$ in $\mathbb{H}$ which is mapped onto itself by a nontrivial transformation $A \in \Gamma'$. A closed geodesic γ is therefore one which can be parameterized by the circle S^1. γ is a *simple geodesic* if for a preimage $\tilde{\gamma}$ in $\mathbb{H}$ and any transformation $A \in \Gamma'$, $\tilde{\gamma} \cap A(\tilde{\gamma}) \neq \phi$ implies that $\tilde{\gamma} = A(\tilde{\gamma})$. Thus γ will have no transverse self intersections, see Figure 4.2.



a. b. c.

Fig. 4.2. (a) A simple closed geodesic. (b) A closed geodesic which is not simple. (c) A simple geodesic which is not closed. One end is asymptotic to the puncture and the other end is asymptotic to a simple closed geodesic.

Due to the simplicity of the surface T_z there is an elegant relationship between generators of Γ' and simple closed geodesics on T_z which was first observed by Nielsen.

Proposition 4.3. [N]. *A geodesic $\tilde{\gamma}$ in $\mathbb{H}$ projects to a simple closed geodesic γ on T_z if and only if $\tilde{\gamma}$ is the fixed axis of a generator of Γ' .*

Combining this proposition with Cohn's Theorem 2.2 we may conclude that

(4.4) *γ on T_z is a simple closed geodesic if and only if γ is the projection of a Markoff geodesic in $\mathbb{H}$.*

We shall refer to a geodesic on T_z which is the projection of a Markoff geodesic in $\mathbb{H}$ also as a *Markoff geodesic*.

It only remains to define "Markoff geodesic" intrinsically in the geometry of T_z and without recourse to the Markoff theorem.

When $k \geq 1$ the half plane $\tilde{\mathcal{U}}_k = \{z \mid \mathrm{Im} z > 1/k\}$ in $\mathbb{H}$ projects to a domain $\mathcal{U}_k$ on the surface T_z which is conformally a punctured disc. $\mathcal{U}_k$ is called a *cusp neighborhood* on T_z and is characterized intrinsically by its hyperbolic area, which is $6k$.

Let us put things together. We just saw that the simple closed geodesics on T_z are precisely the Markoff geodesics. Applying 3.2 we conclude that $\tilde{\gamma}$ in $\mathbb{H}$ covers a simple closed geodesic γ on T_z if and only if $H(\tilde{\gamma}) < 3/2$. The last inequality may be rephrased as follows: for all $A \in \Gamma'$, $A(\tilde{\gamma}) \cap \tilde{\mathcal{U}}_{2/3} = \phi$. Down on T_z this is equivalent to the assertion that γ is disjoint from $\mathcal{U}_{2/3}$, the cusp neighborhood of area 4.

We have proven

Theorem 4.5. *The geodesic γ on T_z is a simple closed geodesic if and only if γ lies in the complement of a cusp neighborhood of area 4.*

This is an intrinsic geometric version of Markoff's theorem. It says that simple closed geodesics stay further away from the puncture than all other geodesics.

The surface T_Z represents only one of many possible conformal structures which can be put on a punctured torus. In fact there is a two real parameter family of such structures. It is natural to ask whether an analogue to Theorem 4.5 holds for these other surfaces. The answer is yes.

Let γ be a geodesic on a hyperbolic once-punctured surface N. The quantity $A(\gamma)$ is defined as the area of the largest open cusp neighborhood on N disjoint from γ. See Figure 4.6.

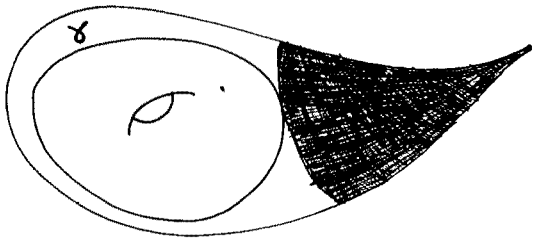


Fig. 4.6. The shaded region represents the largest cusp neighborhood which is disjoint from γ.

The discussion prior to the statement of Theorem 4.5 shows that when $N = T_Z$, $A(\gamma) = 6H^{-1}(\tilde{\gamma})$ where $\tilde{\gamma}$ is a lift of γ to $\mathbb{H}$. Consequently, a form f with $\tilde{\gamma} = \gamma_f$ satisfies $12M(f) = A(\gamma)$. Thus $A(\gamma)$ is the intrinsic analogue to the minimum $M(f)$.

Let S denote the set of simple closed geodesics on a hyperbolic surface N and let $\bar{S}$ be the set of geodesics which are limits of those in S. In other words, a geodesic γ belongs to $\bar{S}$ if there is a sequence γ_n of simple closed geodesics on N with lifts $\tilde{\gamma}_i$ having endpoints x_i and y_i in $\hat{\mathbb{R}}$ so that $\lim_{i\to\infty} x_i = x$ and $\lim_{i\to\infty} y_i = y$ where x and y are the endpoints of a lift $\tilde{\gamma}$ of γ.

All of the geodesics belonging to $\bar{S}$ are simple, but in general there may be simple geodesics with compact support on N which are not in $\bar{S}$. This is the case, for example, when N is a punctured torus.

We can now state 4.5 in a strengthened form.

Theorem 4.7. *A geodesic* γ *on a hyperbolic punctured torus* T *belongs to* $\bar{S}$ *if and only if* $A(\gamma) \geq 4$. *When* γ *belongs to* S, $A(\gamma) = 4 \coth \frac{\ell(\gamma)}{2}$. *When* γ *is a limit geodesic in* $\bar{S} \setminus S$, $A(\gamma) = 4$.

This theorem can be proven directly using geometric and topological arguments as in [H] or it can be derived as a corollary to Schmidt's Theorem 4.1 in [Sc] by translating the results there as we did above for Cohn's version 2.2 of the Markoff theorem [Sh].

Similar results are known for other punctured surfaces: the hyperbolic four times punctured spheres [L-S,H], and the branched surfaces with signature $(0;2,2,2,\infty)$ [Sh]. That we have such knowledge regarding these particular surfaces is explained by the fact that every $(0;2,2,2,\infty)$ surface is finitely covered by a punctured torus as well as by a four times punctured sphere. Moreover, as both Series and Sheingorn have observed the spaces $\bar{S}$ are in a sense the same for all three surfaces.

In general, the structure of the spectrum of values $A(\gamma)$ for geodesics on a surface N, which we call the *Cohn spectrum* of N, will not admit a description of the form of 4.7; although one is led to suspect that a Markoff-like structure for the upper part of the spectrum is not uncommon. To see the difficulty in higher genus we consider the punctured surface N illustrated in Figure 4.8 with the geodesic α separating N into subsurfaces N_1 and N_2. Any geodesic lying on N_1 is closer to the cusp than a geodesic on N_2. Simply knowing whether a geodesic γ on N self intersects will afford little assistance in computing $A(\gamma)$.

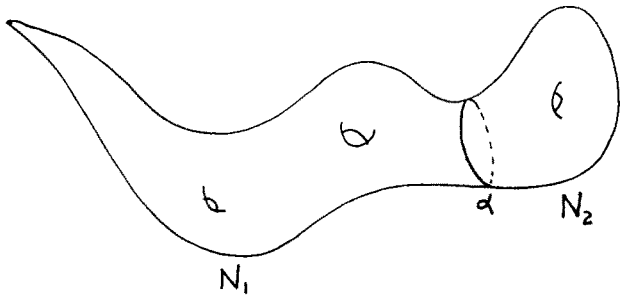


Fig. 4.8. For higher genus surfaces the Cohn spectrum will have a more complex structure.

References

[C1] H. Cohn, Approach to Markoff's minimal forms through modular functions, Annals of Math. 61 (1955) 1-12.

[C2] _____, Representation of Markoff's binary quadratic forms by geodesics on a perforated torus, Acta Arith. 18 (1971) 125-136.

[C3] _____, Markoff forms and primitive words, Math. Annalen, 196 (1972) 8-22.

[C4] _____, Some direct limits of primitive homotopy words and of Markoff geodesics, *Discontinuous Groups and Riemann Surfaces*, Annals Math. Studies, vol. 79, Princeton, 1974, 81-98.

[C5] _____, Minimal geodesics on Fricke's torus-covering, *Riemann Surfaces and Related Topics*, Annals Math. Studies, vol. 97, Princeton, 1980, 73-85.

[H] A. Haas, Diophantine approximation on hyperbolic Riemann surfaces, to appear in Acta Math. 156:1-2.

[L-S] J. Lehner & M. Sheingorn, Simple closed geodesics on $H^+/\Gamma(3)$ arise from the Markov spectrum, Bulletin of the A.M.S., 11 (1984), 359-362.

[M] A. A. Markoff, Sur les formes binaires indefinies, I,*Math. Ann* 15 (1879), 281-309; II, 17 (1880), 379-400.

[N] J. Nielsen, Die isomorphismen der allgemeinen unendlichen gruppe mit zwei erzeugenden, *Math. Ann.*78 (1918), 385-397.

[Sc] A. L Schmidt, Minimum of quadratic forms with respect to Fuchsian groups, I, *J. Reine Angev. Math.* 286/287(1976), 341-368.

[Se] C. Series, The geometry of Markoff numbers, The Math. Intel. 7 (1985).

[Sh] M. Sheingorn, Characterization of simple closed geodesics on Fricke surfaces, Duke Math. Jnl. 52 (1985), 535-545.