Duke’s Theorem and Continued Fractions

John Mangual

February 20, 2008

Abstract

For uniformly chosen random $\alpha \in [0,1]$, it is known the probability
the $n^{\text{th}}$ digit of the continued-fraction expansion, $[\alpha]_n$ converges to
the Gauss-Kuzmin distribution $\mathbb{P}([\alpha]_n = k) \approx \log_2(1 + 1/k(k + 2))$
as $n \to \infty$. In this paper, we show the continued fraction digits of
$\sqrt{d}$, which are eventually periodic, also converge to the Gauss-Kuzmin
distribution as $d \to \infty$ with bounded class number, $h(d)$. The proof
uses properties of the geodesic flow in the unit tangent bundle of the
modular surface, $T^1(\text{SL}_2\mathbb{Z}\backslash \mathbb{H})$.

1 Continued Fractions...

For any $\alpha \in [0,1]$ we can define the continued fraction expansion in
$\mathbb{Z}$ by repeating a two step algorithm. First $a_0 = \alpha$ and $b_0 = [\alpha]$. Now
we simply repeat:

$$a_{k+1} = \{1/a_k\} \quad b_{k+1} = [a_k] \quad (1)$$

The end result is that $\alpha$ can be encoded as a sequence of integers:
$[b_0, b_1, b_2, \ldots]$.

If $\alpha$ is rational then we get a finite continued fraction. What if $\alpha$
is the square root of a irrational number? Then we get an eventually
repeating sequences of numbers $b_k$. For example, the sequence for $\sqrt{7}$
is $[2, 1, 1, 1, 4, 1, 1, 1, 4, \ldots]$ where the $[1, 1, 1, 4]$ motif repeats forever.
How can we get a purely periodic sequence? A theorem by Galois
says:
Theorem 1.1 (5). A quadratic number $\alpha$ has a purely periodic continued fraction expansion if and only if $\alpha > 1$ and $-1 < \alpha' < 0$ where $\alpha, \alpha'$ have the same quadratic equation.

Now let’s ask about statistics of these continued fractions. How often does the number 5 appear in a generic continued fraction? This answer for a random $\alpha \in [0, 1]$ chosen uniformly was found by Kuzmin in 1928. He showed:

Theorem 1.2 (5). There exist positive constants $A, B$ such that

$$\left| A_n(k) - \log_2 \left( 1 + \frac{1}{k(k+2)} \right) \right| \leq \frac{A}{k(k+1)} e^{-B\sqrt{n-1}}$$

Where $A_n(k) = |\{x \in [0, 1] : b_n(x) = k\}|$.

In this paper we look at how the statistics of the continued fraction digits of $\sqrt{d}$ for $d > 0$ behave as $d \to \infty$. In fact we have to be more specific and restrict ourselves to the case of bounded class number, so $h(d)$ is less than some constant. Also, since our sequence $b_k(\sqrt{d})$ is deterministic, we need to define the statistics we’ll be looking at:

$$c(\alpha, k) = \lim_{T \to \infty} \frac{\# \{0 \leq i < T : b_i(\alpha) = k\}}{T}$$

We claim that these statistics approach the limit above, i.e.

Theorem 1.3. As $d \to \infty$ with $h(d)$ bounded:

$$\lim_{d \to \infty} c(\sqrt{d}, k) \to \log_2 \left( 1 + \frac{1}{k(k+2)} \right)$$

To prove this we’re going, as $d \to \infty$ and $h(d) = 1$, the orbits of $\sqrt{d}$ under the map $T : x \mapsto \{1/x\}$ approach the Gauss-Kuzmin on $[0, 1]$. We can rephrase Theorem 1.4 in this new language:

Theorem 1.4. Let $x_0 = \{\sqrt{d}\}$, $h(d) = 1$, $T : x \mapsto \{1/x\}$ be the Gauss map and $f : [0, 1] \to \mathbb{R}$ be continuous:

$$\lim_{d \to \infty} \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f(T^k(x_0)) = \int_0^1 \frac{f(x)}{\ln 2} \cdot \frac{dx}{1 + x}$$

where $h(d) = 1$ as $d$ goes to infinity.
To prove this we need to change settings and examine geodesics in the upper half plane.

2 ... and the Geodesic Flow

Let’s switch contexts to $\mathbb{H} = \{ x + iy : y > 0 \}$ as a differentiable manifold with the Poincaré metric:

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

Much of this exposition follows [2], Chapters 3 and 13.

The geodesics in this metric are (Euclidean) semi-circles with diameters along the real line. Thus for any unit tangent vector $(z, e^{i\theta}) \in T^1(\mathbb{H})$ there is a unique oriented geodesic which goes through $z$ and whose tangent at $z$ points in the direction $e^{i\theta}$.

The group $\text{SL}(2, \mathbb{R})$ acts on $\mathbb{H}$ by fractional linear transformations:

$$ g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : z \mapsto \frac{az + b}{cz + d} $$

Then points in the complex plane are identified as points in the projective complex line $\mathbb{P}^1(\mathbb{C})$:

$$ \left( \begin{array}{c} z \\ 1 \end{array} \right) \mapsto \left( \begin{array}{c} \frac{az + b}{cz + d} \\ 1 \end{array} \right) $$

This is simply the usual matrix action of $\text{PSL}(2, \mathbb{R})$.

We can also consider the quotient group of $\mathbb{H}$ under the action of $\text{SL}_2(\mathbb{Z})$. The quotient under this group action $\text{SL}_2(\mathbb{Z}) \backslash \mathbb{H}$ has is the fundamental domain represented by the intersection of four sets $\{ \text{Im}(z) > 0 \}, \{|z| < 1\}, \{|z + 1| < 1\}$ and $\{|z - 1| < 1\}$.

The tangent space of $\mathbb{H}$ is simply $\mathbb{H} \times \mathbb{C}$. Any element $g \in \text{SL}(2, \mathbb{R})$ can act on the tangent bundle by:

$$ Dg(z, v) = (g(z), g'(z)v) = \left( \frac{az + b}{cz + d}, \frac{v}{(cz + d)^2} \right) $$
It turns out this action is simply transitive and therefore

\[ T^1(\mathbb{H}) \simeq \text{PSL}(2, \mathbb{R}) \]

where \( T^1(\mathbb{H}) \) is the unit tangent bundle.

Now any two points, \( z_1, z_2 \in \mathbb{H} \) determine a unique geodesic - the unique Euclidean semi-circle passing through \( z_1 \) and \( z_2 \). We can consider a map, \( \mathcal{G}_t \), which flows a tangent vector along its geodesic exactly \( t \) units of arc length. This is the geodesic flow from \( z_1 \) to \( z_2 \). You need to specify both a starting point an a direction, an element of \( S^1 \), to get a unique geodesic. Equivalently we can describe the geodesic as right-multiplication by the elements:

\[ a_t = \begin{pmatrix} e^{-t/2} & 0 \\ 0 & e^{t/2} \end{pmatrix} \]

This will be the basis for defining the continued fraction map in terms of the geodesic flow. Furthermore we can define the geodesic flow restricted to \( T^1(\text{SL}_2(\mathbb{Z})) \).

In light of Theorem 1.1, we should consider geodesics whose endpoints \( \alpha, \alpha' \in \mathbb{Q}[\sqrt{d}] \) for some \( d > 0 \) and such that \( \alpha > 1 \) and \( -1 < \alpha' < 0 \). These curves necessarily cut the set \( \{ y_i : 0 < y < 1 \} \) transversely. In fact, we can identify these geodesics either by their endpoints or by the tangent vector at which they cut \([0, i] \).

The elements of \( B = T^1(\{ y_i : 0 < y < 1 \}) \) parameterize these geodesics by the angle at which they cut the line segment \([0, i] \). If we only consider the first case, the set is called \( B^+ \) and in the second case it is called \( B^- \).

\[
B^+ = \left\{ (y_i, e^{i\theta}) : 0 < y < 1, -\frac{\pi}{2} < \theta < 0 \right\} \\
B^- = \left\{ (y_i, e^{i\theta}) : 0 < y < 1, \pi < \theta < \frac{3\pi}{2} \right\}
\]

These correspond to purely periodic continued fractions and therefore to closed geodesics in the Riemann surface \( \text{SL}_2(\mathbb{Z}) \backslash \mathbb{H} \).

**Definition 2.1.** Consider a geodesic, \( \gamma \) in \( \mathbb{H} \) with endpoints \( \alpha, \alpha' \) for which \( \alpha < -1 \) and \( 1 > \alpha' > 0 \). Define **natural coordinates** by \((y, z)\) where \( y = \alpha \) and \( z = \frac{1}{\alpha + \alpha'} \).
**Definition 2.2.** Define \( T : B \to B \) in terms of the geodesic flow \( G_t \) by:

\[
T[(z, e^{i\theta})] = G_{t_0}(z, e^{i\theta}) \quad \text{where} \quad t_0 = \inf \{ t > 0 : G_t(z, e^{i\theta}) \in \text{SL}_2(\mathbb{Z})(B) \}
\]

This \( t_0 \) may not be finite but whenever it is finite this map is well-defined. This is known as the return time map for the cross section \( B \).

**Lemma 2.1.** Let \( x = (ib, e^{i\theta}) \in B_+ \) have natural coordinates \((y, z)\). Then \( T(x) \in \text{SL}_2(\mathbb{Z})(B) \) if \( T \) is defined on \( x \). Moreover \( T(x) \in \text{SL}_2(\mathbb{Z})(x') \), where \( x' \) has natural coordinates

\[
\overline{T}(y, z) = \left( \begin{array}{c} 1 \\ y \end{array} \right), y(1 - yz)
\]

A similar property holds for \( B_- \).

The first step in the proof of Theorem 1.3 is to consider closed geodesics in \( \text{SL}_2(\mathbb{Z}) \backslash \mathbb{H} \) of discriminant \( d \). They correspond to purely periodic continued fractions and to elements of \( B \). Their natural coordinates lie in the ring \( \mathbb{Q}[\sqrt{d}] \). Thus, we can prove these geodesics become equidistributed in \( B \) as \( d \to \infty \) and \( h(d) \leq M \), then the first natural coordinate follows the Gauss-Kuzmin distribution in \([0, 1]\). In other words, the orbit under \( T : x \mapsto \{1/x\} \) of any element of \( \mathbb{Q}[\sqrt{d}] \) becomes Gauss-Kuzmin in \([0, 1]\), asymptotically.

### 3 Duke’s Theorem

We to define some special sets of geodesics:

**Definition 3.1.** Let \( d < 0 \) and let \((x_d, y_d)\) be the fundamental solution to \( x^2 - dy^2 = 4 \). Define \( \Gamma_d \) as the set of geodesics in \( \text{PSL}_2(\mathbb{Z}) \backslash \mathbb{H} \) of length \( d \) induced by quadratic forms \( q(x, y) = ax^2 + bxy + cy^2 \) with \( b^2 - 4ac = d \).

We can consider closed geodesics on \( \text{SL}_2(\mathbb{Z}) \backslash \mathbb{H} \)

**Theorem 3.1** (1). Suppose \( d \) is a fundamental discriminant. Then for some \( \delta > 0 \) depending only on \( \Omega \)

\[
\frac{\sum_{C \in \Lambda_d} |C \cap \Omega|}{\sum_{C \in \Lambda_d} |C|} = \mu(\Omega) + O(d^{-\delta}) \quad (2)
\]
as \( d \to \infty \) where \(|C|\) is the non-Euclidean length of \( C \) and the \( \mathcal{O} \) constant depends only on \( \delta \) and \( \Omega \).

The idea that geodesic orbits in homogeneous spaces become equidistributed can be extended to the tangent bundle. In this case, we consider the geodesic flow in \( T^1\text{SL}_2(\mathbb{Z}) \). This is useful because the set we wish to consider \( B, B^+ \), and \( B^- \), who live in the unit tangent bundle and not the underlying space. Fortunately for us, Duke’s theorem extends to this case well.

**Definition 3.2.** Define \( \Lambda_d = \{ \gamma[q] : [q] \in \text{PSL}_2(\mathbb{Z}) \setminus Q_d(\mathbb{Z}) \} \), the geodesics associated with the set of binary quadratic forms modulo \( \text{PSL}_2(\mathbb{Z}) \) equivalence. If \( q(x, y) = ax^2 + bxy + cy^2 \), the geodesic \( \gamma[q] \) has endpoints defined by \( q(x, 1) = 0 \). Then project this geodesic onto the modular surface \( \text{PSL}_2(\mathbb{Z}) \setminus \mathbb{H} \).

**Theorem 3.2 (4).** As \( d \to \infty \), \( d \equiv 0, 1 \mod 4 \), \( d \) not a perfect square, the set \( \Gamma_d \) becomes equidistributed on the unit tangent bundle, \( T^1\text{SL}_2(\mathbb{Z}) \), with respect to the volume measure \( d\mu_L = \frac{3\pi}{4} dy dx y^2 \).

\[
\sum_{C \in \Gamma_d} |C \cap \Omega| \quad \sum_{C \in \Gamma_d} |C| = \mu_L(\Omega) + \mathcal{O}(d^{-\delta})
\]

Now we are ready to prove our equidistribution result. Theorem 1.4 can be rephrased in dynamical into ergodic theory language. Proving it for \( f(x) = \chi_I(x) \) for some interval \( I \in [0, 1] \), we can prove it for any continuous function \( f(x) \).

**Theorem 3.3.** Let \( T : [0, 1] \to [0, 1] \) be the map defined in Theorem 2.1. Then the orbit \( \{ T^n(x) : n \in \mathbb{N} \} \) is distributed as the Gauss-Kuzmin distribution as \( d \) goes to infinity. Specifically, for any interval \( I \subseteq [0, 1] \):

\[
\lim_{d \to \infty} \frac{\# \{0 \leq n < l(d) : T^n(B) \in I \}}{l(d)} = \frac{1}{\ln 2} \int_I \frac{dx}{1 + x}
\]

Where \( l(d) \) is the period of the continued fraction corresponding to \( \sqrt{d} \).

**Proof.** Let \( I \subseteq [0, 1] \) be an interval and \((y, z)\) be the natural coordinates in \( B \).

\[
B_{I, \epsilon} = \{(y, z) \in B : y \in I \} \times [-\epsilon/2, \epsilon/2] 
\]
This set can be embedded in $T^1(\text{SL}_2(\mathbb{Z}) \backslash \mathbb{H})$ as:

$$\{ \mathcal{G}_t(B) : t \in [-\epsilon/2, \epsilon/2] \}$$

The Haar measure in the natural coordinates is Lebesgue in all three variables.

By Duke’s theorem, there is $\delta > 0$ such that

$$\sum_{C \in \Gamma_D} |C \cap B_{I,\epsilon}| = \mu(B_{I,\epsilon}) + O(d^{-\delta})$$

Note that this equation is true even if $I = [0, 1]$. It therefore follows that:

$$\frac{\sum_{C \in \Gamma_D} |C \cap B_{I,\epsilon}|}{\sum_{C \in \Gamma_D} |C|} = \frac{\mu(B_{I,\epsilon})}{\mu(B_{[0,1],\epsilon})} + O(d^{-\delta})$$

Because of how we defined $B_{I,\epsilon}$, the total length is simply the number of times the geodesics $\Gamma_d$ cut $I$ times $\epsilon$:

$$\sum_{C \in \Gamma_D} |C \cap B_{I,\epsilon}| = \epsilon \cdot \sum_{C \in \Gamma_D} \# \{ C \cap I \}$$

$$= \epsilon \cdot \# \{ 0 \leq n < l(d) : T^n(x_0) \in I \}$$

Here $l(d)$ denotes the period of the continued fraction with respect to $l(d)$. In the case $I = [0, 1]$ the last line is just $\epsilon \cdot l(d)$ so the left hand side of (4) is really just counting measure:

$$\frac{\# \{ 0 \leq n < l(d) : T^n(x_0) \in I \}}{l(d)}$$

For the right hand side of (4) let’s find the measure of $B_{I,\epsilon}$:

$$\mu(B_{I,\epsilon}) = \epsilon \int_I \int_0^{1+y} \frac{dydz}{\ln 2} = \frac{\epsilon}{\ln 2} \int_I \frac{dy}{1 + y}$$

Therefore equation (4) should read:

$$\lim_{d \to \infty} \frac{\# \{ 0 \leq n < l(d) : T^n(x_0) \in I \}}{l(d)} = \frac{1}{\ln 2} \int_I \frac{dy}{1 + y}$$
4 Bounded Class Number

We can relax the condition $h(d) = 1$ in Theorem 1.4:

**Theorem 4.1.** Let $x_0 = \{\sqrt{d}\}$, $h(d) = 1$, $T : x \mapsto \{1/x\}$ be the Gauss map and $f : [0,1] \rightarrow \mathbb{R}$ be continuous:

$$\lim_{d \rightarrow \infty} \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} f(T^k(x_0)) = \int_0^1 \frac{f(x)}{\ln 2} \cdot \frac{dx}{1 + x}$$

where there exists $M > 0$ such that $h(d) < M$.

**Proof.** We simply examine Theorem 3.3 more closely. In this case, $\Gamma_d$ has more than one element, i.e. there are several geodesics with the same discriminant $d$. Duke’s theorem says the sum

$$\sum_{C \in \Lambda_d} |C \cap \Omega| = \sum_{C \in \Lambda_d} |C \cap \Omega| = \mu(\Omega) + O(d^{-\delta})$$

is the Lebesgue measure. Here $L_d$ is the length of a geodesic of discriminant $d$. Following the principle outlined in [6] (Section 1.3.5 (1)), since the Lebesgue measure is an extreme point in the convex space of geodesic-flow invariant measures on $T^1(\text{SL}_2(\mathbb{Z}))$, since the Gauss map $T$ is ergodic and since their sum is a Lebesgue measure, each term in the sum must also approach Lebesgue measure. \(\square\)

Then continued fractions in $\mathbb{Q}[\sqrt{d}]$ tend towards the Gauss Kuzmin distribution even in the case bounded class number. This behavior is truly generic.

**References**

[1] William Duke. “Hyperbolic distribution problems and half-integral weight Maass forms”, Invent. Math. 92 (1988), no. 1, 73-90.

[2] Manfred Einsiedler, Thomas Ward. “Ergodic Theory with a View Towards Number Theory” (draft). [http://www.mth.uea.ac.uk/ergodic/](http://www.mth.uea.ac.uk/ergodic/)

[3] Svetlana Katok, Ilie Ugarcovici “Symbolic dynamics for the modular surface and beyond” Bull. of the Amer. Math. Soc., 44, no. 1 (2007), 87-132
[4] Phillipe Michel, Akshay Venkatesh. “Equidistribution, L-functions and Ergodic Theory: on some problems of Yu. Linnik.” International Congress of Mathematicians 2006, Madrid, Volume II, 421-458.

[5] Stephen Miller, Ramin Takloo-Bighash. “An Invitation to Modern Number Theory”, Princeton University Press, 2006.

[6] Akshay Venkatesh. “Sparse equidistribution problems, period bounds, and subconvexity” arXiv:math/0506224