Continued fractions from Euclid to the present day

Philippe Flajolet, Brigitte Vallée, and Ilan Vardi

Contents

It may appear surprising that the following books have an important feature in common:

- G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers*, Clarendon Press, Oxford, 1979.
- D.H. Fowler, *The Mathematics of Plato's Academy: a New Reconstruction*, Clarendon Press, Oxford 1990.
- D.E. Knuth, *The Art of Computer Programming, Vol. 2*, Addison–Wesley 1981.
- P. Billingsley, *Ergodic Theory and Information*, Wiley, New York 1965.

Indeed, each of these books presents an excellent introduction to the theory of continued fractions. Of course, the points of view are different, and it is the goal of this paper to indicate how continued fractions are relevant to

- Number theory.
- Ancient Greek mathematics.
- Analysis of algorithms.
- Probability theory.

In the past decade, two fairly sophisticated techniques have been applied to the analysis of the running time of the Euclidean algorithm. On the one hand, the application of transfer operators by Brigitte Vallée, on the other, the application of modular forms by Ilan Vardi.

Thus, as well as being a survey of of the theory of continued fractions, this paper is meant to serve as an introduction and description of these methods and their application.

1 Examples

Continued fractions are a generalization of compound fractions like $\frac{14}{11} = 1\frac{3}{11}$. For example,

$$\frac{14}{11} = 1\frac{3}{11} = 1 + \frac{1}{11/3} = 1 + \frac{1}{3\frac{2}{3}} = 1 + \cfrac{1}{3 + \cfrac{1}{3/2}} = 1 + \cfrac{1}{3 + \cfrac{1}{1\frac{1}{2}}} = 1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{2}}},$$

(if you're clever, you can actually do one more step). This works for every rational number p/q which can be written in the form

$$\frac{p}{q} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_r}}}$$

where $a_1, \ldots, a_r$ are positive integers. For ease of notation this is usually written as $\frac{p}{q} = [a_0, a_1, \ldots, a_r]$. This works just as well for real numbers, for example,

$$\pi = 3 + .141592653\ldots = 3 + \cfrac{1}{7 + .062513305\ldots} = 3 + \cfrac{1}{7 + \cfrac{1}{15 + .99659440\ldots}}$$

$$= 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + .003417231\ldots}}} = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292 + .63459088\ldots}}}}$$

but only terminates for rational numbers. In general, every real number x has an expansion

$$x = [a_0(x), a_1(x), a_2(x), \ldots]$$

It is interesting to experiment with different numbers and find patterns:

$$\begin{aligned}
\sqrt{7} &= [2,1,1,1,4,1,1,1,4,1,1,1,4,1,1,1,4,1,1,1,4,1,1,1,4,\ldots], \\
e &= [2,1,2,1,1,4,1,1,6,1,1,8,1,1,10,1,1,12,1,1,14,1,1,16,\ldots], \\
\pi &= [3,7,15,1,292,1,1,1,2,1,3,1,14,2,1,1,2,2,2,2,1,84,\ldots].
\end{aligned}$$

The patterns seen in these expansions answer the often asked question as to why people compute lots of digits of π, and not of $\sqrt{2}$ or e–they already have a known continued fraction expansion. Thus, one can argue that continued fraction records are more natural. Even though more than a billion base 10 digits of π are known, the record for continued fraction digits is still 17 million as computed by R.W. Gosper in 1985. This computation is equivalent to computing about 18 million base 10 digits, as will be seen below.

The most important property of the continued fraction expansion is that the rational number you get from taking an initial segment of the expansion, usually called a *convergent*, gives a record breaking approximations to your number. In other words, if $x = [a_0, a_1, a_2, \ldots, a_n, a_{n+1}, \ldots]$, then the nth convergent $\frac{p}{q} = [a_0, \ldots, a_n]$ and will be closer to x than any $\frac{p'}{q'}$ for $q' < q$.

This property is especially interesting for approximations to square roots of integers. Thus, the convergents p/q to $\sqrt{D}$ are exactly the solutions of Pell's equation $p^2 - Dq^2 = 1$. It seems that this equation might first have appeared in Archimedes' Cattle Problem see [85].

Exercise. Characterize all the record breaking approximations to a real number. In particular, do they all come from cutting off a continued fraction expansion?

As an example of the above process, $[3, 7] = \dfrac{22}{7}$ gives the best approximation to π with denominator less than or equal to 7. Moreover, cutting off the expansion just before a large digit gives an especially good approximation. So to get a very good approximation to π, take the digits before the 292 and rationalize

$$[3, 7, 15, 1] = \frac{355}{113} = 3.14159292\ldots$$

is a good approximation to $\pi = 3.141592653\ldots$. It will be seen below that in fact, the increase in accuracy is in fact proportional to the next coefficient.

Exercise. Archimedes [2] proved the bounds $3 + \dfrac{10}{71} < \pi < 3 + \dfrac{1}{7}$. What does this say about the continued fraction expansion of π?

2 Culture

Continued fractions have fascinated mankind for centuries if not millennia. The timeless construction of a rectangle obeying the "divine proportion" (the term is in fact from the Renaissance) and the "self-similarity" properties that go along with it are nothing but geometric counterparts of the continued fraction expansion of the golden ratio,

$$\phi \equiv \frac{1+\sqrt{5}}{2} = \cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\ddots}}}.$$

Geometry in India developed from the rules for the construction of altars. The *Sulva Sūtra* (a part of the *Kalpa Sūtra* hypothesized to have been written around 800 BC) provides a rule[1] for doubling an area that corresponds to the near-equality:

$$\sqrt{2} \doteq 1 + \frac{1}{3} + \frac{1}{3\times 4} - \frac{1}{3\times 4\times 34} \qquad \text{(correct to } 2\times 10^{-6}\text{)}. \tag{1}$$

Note that the third and fourth partial sums in (1), namely $\frac{17}{12}$ and $\frac{577}{408}$, are respectively the fourth and eighth convergents to $\sqrt{2}$.

Accordingly, in the classical Greek world, there is evidence of knowledge of the continued fraction for $\sqrt{2}$ which appears in the works of Theon of Smyrna (discussed in Fowler's reconstruction [29] and in [86]) and possibly of Plato in *Theaetetus*, see [8].

Such concerns are still relevant to modern man. In particular, the French printout of this paper uses the paper size A4, which was recently chosen due to its self–similar properties. In particular, it has approximately the same proportion when folded in half. By definition, the dimensions of A4 paper are length = 29.7 centimeters and width = 21 centimeters. The proportion is a convergent to $\sqrt{2}$ since

$$\frac{29.7}{21} = \frac{99}{70} = [1,2,2,2,2,2], \quad \text{while} \quad \sqrt{2} = [1,2,2,\ldots].$$

(Note that $(1+\sqrt{2})^6 = 99 + 70\sqrt{2}$.) Moreover, $16\,(.21\times .297) = .99792$, so that there are almost exactly 16 sheets of A4 paper in a square meter.

Perhaps the most important application of continued fractions is to the theory of calendars. Thus, in the western calendar, the years keep track of the seasons, but months serve absolutely no purpose (in fact, as has been noted by fans of the National Football League, it is much more useful to keep track of weeks). Thus, March will be Spring in the Northern hemisphere, but knowing that it is March 15 does not provide any extra information.

On the other hand, the Islamic calendar is a purely lunar calendar, i.e., dates of the month correspond to phases of the moon, but since the year has 354 days, this calendar does not keep track of seasons. Thus, there is a full moon on the 15th of Ramadan, but it is not clear what season it is unless the year is specified.

To be a true representation of nature, a calendar which uses years and months should keep track of seasons and phases of the moon. In order to do this, one looks at the fraction 365.24/29.53, the ratio of the solar year to the lunar month. The continued fraction expansion is

$$\frac{365.24}{29.53} = [12,2,1,2,2,155]\,.$$

[1] "Increase the measure by its third part, and this third part by its own fourth, less the thirty-fourth part of that fourth". See vol. I of Dutt's book [23, p. 272] for context including otherwise rational approximations to $\sqrt{\pi/4}$.

Cutting this off before the very large digit 155 gives $[12,2,1,2,2] = \frac{235}{19} = 12\,\frac{7}{19}$. This says that there are almost exactly 235 lunar months in 19 years. The corresponding Metonic cycle of 7 leap months every 19 years is therefore best way to keep a lunar and solar calendar (the western calendar is purely solar). This method is used in the Jewish calendar with the extra month being Adar II, see [18]. Thus, the 15th of Nissan (=Passover) always occurs in Spring and during the full moon.

3 Euclid

It is clear that the process involved in obtaining the continued fraction expansion of the rational number p/q is simply the Euclidean algorithm to find the greatest common divisor of the two integers p and q. From this point of view, one can consider the continued fraction expansion to be a record of a Euclidean algorithm. In general, one can expect continued fraction coefficients whenever a Euclidean algorithm is used.

The Euclidean algorithm to compute $\mathrm{GCD}\,(p,q)$, the greatest common divisor of two numbers, can be given by the following recursion

$$\mathrm{GCD}\,(p,q) = \mathrm{GCD}\,(p \bmod q, q), \qquad \mathrm{GCD}\,(p,q) = \mathrm{GCD}\,(q,p)\,.$$

One can therefore expect that a function $F(p,q)$ satisfying

$$F(p,q) = F(p \bmod q, q), \qquad F(p,q) = G(p,q,F(q,p))\,,$$

where G is a simple expression, should be expressible in terms of continued fractions.

For example, the Lagrange symbol $\left(\frac{d}{c}\right)$, which essentially expresses whether d is a perfect square modulo c can be evaluated using the Euclidean reduction

$$\left(\frac{d}{c}\right) = \left(\frac{d \bmod c}{c}\right), \quad \left(\frac{c}{d}\right) = (-1)^{(c-1)(d-1)/4}\left(\frac{d}{c}\right),$$

where c, d are odd (the second identity is quadratic reciprocity). In fact, it was shown by Rademacher [63] that

$$\left(\frac{d}{c}\right) = (-1)^{(3-a-d+c\Sigma(-1)^i a_i)/4}, \quad \text{where } \frac{d}{c} = [0, a_1, \ldots, a_r]\,,$$

and $0 < a, d < c$, $ad \equiv 1 \pmod{c}$ (note that $d^{-1} \bmod c$ can be computed using the Euclidean algorithm), and c and r are odd.

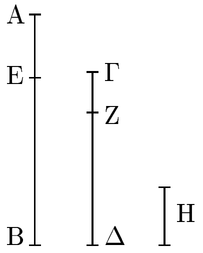
Remark. The identity $n + 1 = n + \frac{1}{1}$ implies that $[a_0, \ldots, a_n] = [a_0, \ldots, a_n - 1, 1]$ if $a_n > 1$. It follows that every expansion can be made to have even or odd length, and alternatively, one can always force an expansion to end with a one or with a number greater than one.

Reciprocally, continued fractions have applications to the Euclidean algorithm. Thus, the length of the continued fraction expansion of p/q is the number of division steps in the Euclidean algorithm applied to p, q. The statistical theory of continued fractions, as will be seen below, will solve the problem of analyzing the (average and typical) running time of the Euclidean algorithm. This explains why a large section of Knuth's "Art of Computer Programming" is devoted to continued fractions. It provides a complete analysis of the running time of the earliest recorded non trivial algorithm. Another excellent survey of the historical aspects of the analysis of the Euclidean algorithm has been given by G. Shallit [76].

It is instructive to actually look up Euclid's description of his algorithm. This is best illustrated by Proposition II of Book VII:

β′

Δύο ἀριθμῶν δοθέντων μὴ πρώτων πρὸς ἀλλήλους τὸ μέγιστον αὐτῶν κοινὸν μέτρον εὑρεῖν.



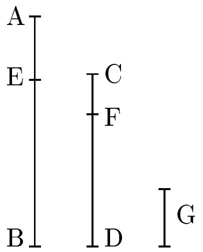
Ἔστωσαν οἱ δοθέντες δύο ἀριθμοὶ μὴ πρῶτοι πρὸς ἀλλήλους οἱ AB, ΓΔ. δεῖ δὴ τῶν AB, ΓΔ τὸ μέγιστον κοινὸν μέτρον εὑρεῖν.

Εἰ μὲν οὖν ὁ ΓΔ τὸν AB μετρεῖ, μετρεῖ δὲ καὶ ἑαυτόν, ὁ ΓΔ ἄρα τῶν ΓΔ, AB κοινὸν μέτρον ἐστίν. καὶ φανερόν, ὅτι καὶ μέγιστον οὐδεὶς γὰρ μείζων τοῦ ΓΔ τὸν ΓΔ μετρήσει.

Εἰ δὲ οὐ μετρεῖ ὁ ΓΔ τὸν AB, τῶν AB, ΓΔ ἀνθυφαιρουμένον ἀεὶ τοῦ ἐλάσσονος ἀπὸ τοῦ μείζονος λειφήσεταί τις ἀριθμός, ὃς μετρήσει τὸν πρὸ ἑαυτοῦ. μονὰς μὲν γὰρ οὐ λειφθήσεται· εἰ δὲ μή, ἔσονται οἱ AB, ΓΔ πρῶτοι πρὸς ἀλλήλους· ὅπερ οὐχ ὑπόκειται. λειφθήσεταί τις ἄρα ἀριθμός, ὃς μετρήσνι τὸν πρὸ ἑαυτοῦ. καὶ ὁ μὲν ΓΔ τὸν BE μετρῶν λειπέτω ἑαυτοῦ ἐλάσσονα τὸν EA, ὁ δὲ EA τὸν ΔZ μετρῶν λειπέτω ἑαυτοῦ ἐλάσσονα τὸν ZΓ, ὁ δὲ ΓZ τὸν AE μετρείτω. ἐπεὶ οὖν ὁ ΓZ τὸν AE μετρεῖ, ὁ δὲ AE τὸν ΔZ μετρεῖ, καὶ ὁ ΓZ ἄρα τὸν ΔZ μετρήσει· μετρεῖ δὲ καὶ ἑαυτόν· καὶ ὅλον ἄρα τὸν ΓΔ μετρήσει. ὁ δὲ ΓΔ τὸν BE μετρεῖ· καὶ ὁ ΓZ ἄρα τὸν BE μετρεῖ· μετρεῖ δὲ καὶ τὸν EA· καὶ ὅλον ἄρα τὸν BA μετρήσει· μετρεῖ δὲ καὶ τὸν ΓΔ· ὁ ΓZἄρα τοὺς AB, ΓΔ μετρεῖ. ὁ ΓZ ἄρα τῶν AB, ΓΔ κοινὸν μέτρον ἐστίν. λέγω δή, ὅτι καὶ μέγιστον. εἰ γὰρ μή ἐστιν ὁ ΓZ τῶν AB, ΓΔ μέγιστον κοινὸν μέτρον, μετρήσει τις τοὺς AB, ΓΔ ἀριθμοὺς ἀριθμὸς μείζων ὢν τοῦ ΓZ. μετρείτω, καὶ ἔστω ὁ H. καὶ ἐπεὶ ὁ H τὸν ΓΔ μετρεῖ, ὁ δὲ ΓΔ τὸν BE μετρεῖ, καὶ ὁ H ἄρα τὸν BE μετρεῖ· μετρεῖ δὲ καὶ ὅλον τὸν BA· καὶ λοιπὸν ἄρα τὸν AE μετρήσει. ὁ δὲ AE τὸν ΔZ μετρεῖ· καὶ ὁ H ἄρα τὸν ΔZ μετρήσει· μετρεῖ δὲ καὶ ὅλον τὸν ΔΓ· καὶ λοιπὸν ἄρα τὸν ΓZ μετρήσει ὁ μείζων τὸν ἐλάσσονα· ὅπερ ἐστὶν ἀδύνατον· οὐκ ἄρα τοὺς AB, ΓΔ ἀριθμοὺς ἀριθμός τις μετρήσει μείζων ὢν τοῦ ΓZ· ὁ ΓZ ἄρα τῶν AB, ΓΔ μέγιστόν ἐστι κοινὸν μέτρον [ὅπερ ἔδει δεῖξαι].

Proposition 2

Given two numbers not prime to one another, to find their greatest common measure.



Let AB, CD be the two given numbers not prime to one another. Thus it is required to find the greatest common measure of AB, CD.

If now CD measures AB–and it also measures itself–CD is a common measure of CD, AB. And it is manifest that it is also the greatest; for no greater number than CD will measure CD.

But, if CD does not measure AB, then, the less of the numbers AB, CD being continually subtracted from the greater, some number will be left which will measure the one before it. For an unit will not be left; otherwise AB, CD will be prime to one another [VII. I], which is contrary to the hypothesis. Therefore some number will be left which will measure the one before it. Now let CD, measuring BE, leave EA less than itself, let EA, measuring DF, leave FC less than itself, and let CF measure AE. Since then, CF measures AE, and AE measures DF, therefore CF will also measure DF. But it also measures itself; therefore it will also measure the whole CD. But CD measures BE. But is also measures EA; therefore it will also measure the whole BA. But it also measures CD; therefore CF measures AB, CD. Therefore CF is a common measure of AB, CD. I say next that it is also the greatest. For, if CF is not the greatest common measure of AB, CD, some number which is greater than CF will measure the numbers AB, CD. Let such a number measure them, and let if be G. Now, since G measures CD, while CD measures BE, G also measures BE. But is also measures the whole BA; therefore it will also measure the remainder AE. But AE measures DF; therefore G will also measure DF. But it also measures the whole DC; therefore it will also measure the remainder CF, that is, the greater will measure the less: which is impossible. Therefore no number which is greater than CF will measure the numbers AB, CD; therefore CF is the greatest common measure of AB, CD. Q. E. D.

The original text is essential for a correct analysis, since modern assumptions are inevitable in the translated text (though Heath's translation used here [27, Vol. 2] is quite faithful). Moreover, the Greek text is difficult to find in the United States, for example, the definitive edition [26] is in Greek only and the French dual language edition [25, Vol. 2] is out of print.

Two aspects of Euclid's argument are striking: (a) He only goes through three iterations and yet states that this is a general proof, (b) To denote the Euclidean algorithm, he uses the (nowhere defined) term ἀνθυφαιρουμένον = "anthuphairoumenon", best translated as "continued alternating subtraction" [8, p. 111].

In fact, (a) is typical of ancient Greek mathematics in that they did not have the modern concept "..." and thus, not unlike modern undergraduates, they wrote mathematical induction arguments only the number of steps required to explain all the necessary ideas (thus Euclid's proof of the infinity of primes only shows that given 3 primes, there is a 4th prime unequal to them).

(b) is much more obscure and in fact, is the main subject of the interesting book of David Fowler [29]. Fowler's thesis is that ancient Greeks from the time of Plato were used to dealing with continued fractions and that they played a central role in their arithmetic. Whether one agrees with this or not, it is a fact that Euclid does not attempt to define "repeated alternating subtraction," i.e., the Euclidean algorithm, whereas he is usually very careful about not having undefined terms. This indicates a great familiarity with this procedure.

Exercise. Show that $\mathrm{GCD}\left(\dfrac{X^i-1}{X-1}, \dfrac{X^j-1}{X-1}\right) = \dfrac{X^{\mathrm{GCD}\,(i,j)}-1}{X-1}$.

4 Matrices

Consider a matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with integer entries and of determinant one. Then

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^{-p} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a-pc & b-pd \\ c & d \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^{-q} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ c-qa & d-qb \end{pmatrix},$$

is simply the Euclidean algorithm performed on b, d and a, c. For example,

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}^{-3} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 9 & 14 \\ 7 & 11 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

corresponds to the Euclidean algorithm performed on 14 and 11. Letting

$$U = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad L = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

then every matrix A of determinant one with non negative integer coefficients will have a unique factorization $A = U^{\alpha_0} L^{\alpha_1} \cdots U^{\alpha_r}$ corresponding to a continued fraction expansion. In the case when $b > a > 0$, $d > c \geq 0$, then the above process gives

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = U^{a_0} L^{a_1} \cdots U^{a_r},$$

where $b/d = [a_0, a_1, \ldots, a_r]$ (so r is even). The above example was

$$\begin{pmatrix} 9 & 14 \\ 7 & 11 \end{pmatrix} = UL^3ULU.$$

In order to see why this is, note that under the above assumptions, $a - pc > 0$ if and only if $b - pd \geq 0$, and $c - qa \geq 0$ if and only if $d - qb \geq 0$.

Exercise. Show that if $b, d > 0$ are relatively prime, then there is a unique matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with determinant one and $b > a > 0$, $d > c \geq 0$. Explain why in this case, r must be even in the corresponding matrix factorization.

Exercise. If $p_n/q_n = [a_0, \ldots, a_n]$, then

$$U^{a_0} L^{a_1} \cdots U^{a_r} = \begin{pmatrix} p_{r-1} & p_r \\ q_{r-1} & q_r \end{pmatrix}, \text{ if } r \text{ is even}, \quad U^{a_0} L^{a_1} \cdots L^{a_r} = \begin{pmatrix} p_r & p_{r-1} \\ q_r & q_{r-1} \end{pmatrix}, \text{ if } r \text{ is odd.}$$

This has immediate applications. For example, one can now explain what happens when one reverses continued fraction coefficients. More precisely, one can find the fraction $[0, a_r, \ldots, a_1]$ given $\frac{b}{d} = [0, a_1, \ldots, a_r]$. First, assume that r is even. Then, by the above, one has

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = L^{a_1} \cdots U^{a_r},$$

where $ad - bc = 1$ and $0 < a, b < d$. It follows that the transpose is

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix} = L^{a_r} U^{a_{r-1}} \cdots U^{a_1},$$

and so $[0, a_r, \ldots, a_1] = \frac{c}{d}$. Since $ad - bc = 1$ and $0 < c < d$, it follows that the reverse continued fraction of b/d is $(-b^{-1} \bmod d)/d$.

Exercise. Show that if r is odd, then the reverse continued fraction of b/d, $0 < b < d$ is $(b^{-1} \bmod d)/d$.

These properties were used by H.J. Smith in 1855 [77] to give a new proof of Fermat's theorem that every prime of the form $4k+1$ is a sum of two squares. The exposition follows van de Poorten [62]: First, consider the set

$$S = \left\{ \frac{2}{p}, \ldots, \frac{(p-1)/2}{p} \right\},$$

and for each member, write its continued fraction expansion such that its last digit is ≥ 2. It follows that the map reversing the digits preserves this set. Since $p \equiv 1 \pmod 4$, the set S has an odd number of elements, and since inversion is an involution (its square is the identity) there must be a fixed point, say x/p.

Now, if the continued fraction expansion of x/p had even length (i.e., r odd), then $x \equiv x^{-1} \pmod p$, and so $x \equiv \pm 1 \pmod p$, which is not possible, as neither $1/p$ nor $(p-1)/p$ belongs to S. It follows that r is even, and by the above, $x^2 \equiv -1 \pmod p$.

Now write $\frac{x}{p} = [0, a_1, \ldots, a_m, a_m, \ldots, a_1]$ and consider the matrix M of determinant one

$$M = \begin{pmatrix} * & * \\ u & v \end{pmatrix} = L^{a_1} U^{a_2} \cdots (U \text{ or } L)^{a_m},$$

Then the matrix corresponding to x/p is the product $M\, M^t$ so that

$$\begin{pmatrix} * & x \\ * & p \end{pmatrix} = M M^t = \begin{pmatrix} * & * \\ u & v \end{pmatrix} \begin{pmatrix} * & u \\ * & v \end{pmatrix} = \begin{pmatrix} * & * \\ * & u^2 + v^2 \end{pmatrix}$$

and $u^2+v^2=p$. Smith did not realize that this gives an efficient algorithm for computing u,v. However, D.E. Knuth has remarked that this appears in an 1848 paper of Serret and Hermite [75] [50, p. 579]. One therefore gets the following algorithm (generalized by Cornachia in 1906 [9, p. 34]):

Algorithm for computing $p = u^2+v^2$:

(a) Find an $x < p/2$ such that $x^2 \equiv -1 \pmod p$.

(b) Perform the Euclidean algorithm on (x,p) until a remainder is $< \sqrt{p}$, then do one more step. The last two remainders are u and v.

Remark. In practice, one can compute part (a) quickly by finding a non–residue modulo p. Thus, if y is a non–residue, then by Euler's theorem $y^{(p-1)/2} \equiv -1 \pmod p$, and so $x \equiv y^{(p-1)/4} \pmod p$ will have the required property. Note that exponentiation modulo p can be done efficiently using repeated squaring.

In theory, one can compute part (a) in polynomial time by using an algorithm of Schoof [70]. In fact, for any *fixed* a, computing the square root of a mod p can done in polynomial time, by computing the number of points on a corresponding elliptic curve, the number of points on an elliptic curve modulo p being polynomial, by Schoof's algorithm.

5 The non Euclidean algorithm

Another possible way to expand as a continued fraction is to consider to take the continued fraction by "excess"

$$\frac{d}{c} = b_0 - \cfrac{1}{b_1 - \cfrac{1}{b_2 - \cfrac{1}{b_3 - \ddots}}},$$

as was considered by D. Zagier [91], D.E. Knuth [49], and by Kirby and Melvin [47]. For example

$$\frac{14}{11} = 2 - \cfrac{1}{2 - \cfrac{1}{2 - \cfrac{1}{3 - \cfrac{1}{2}}}}$$

Note that there are a lot of 2's.

This algorithm also corresponds to a matrix factorization, but instead of L use $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ (these are the standard generators for $SL(2,\mathbf{Z})$, e.g., see [74]). The relation with the other matrix factorization is that $L = USU$. So to generate the matrix factorization in terms of U and S, find the continued fraction digits of $\frac{b}{d} = [a_0, a_1, \ldots]$ and write as $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = U^{a_0} L^{a_1} \cdots$ as above and replace L with USU. For example

$$\begin{pmatrix} 9 & 14 \\ 7 & 11 \end{pmatrix} = UL^3ULU = U(USU)(USU)(USU)U(USU)U = U^2SU^2SU^2SU^3SU^2\,.$$

The presence of many 2's can now be explained: The number of 2's in this expansion essentially corresponds to the *sum* of the odd continued fraction coefficients of b/d.

6 Linear fractional transformations

It is well known that there is a correspondence between matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ and forms $\frac{ax+b}{cx+d}$. This is a group action in the sense that if $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) z = \frac{ax+b}{cx+d}$, then $(AB)z = A(B(z))$ for any two matrices A, B of determinant one. It follows that the above matrix decomposition leads to a continued fraction expansion for $\frac{ax+b}{cx+d}$

$$\frac{9x+14}{7x+11} = 1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1+x}}}}$$

And the continued fraction for $14/11$ is recovered by letting $x = 0$. This correspondence is seen by the fact that

$$\begin{pmatrix} 1 & p \\ 0 & 1 \end{pmatrix} x = p + x, \qquad \begin{pmatrix} 1 & 0 \\ q & 1 \end{pmatrix} x = \cfrac{1}{q + \cfrac{1}{x}}\,.$$

It follows that the expansion $[a_0, \ldots, a_r]$ corresponds to $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = U^{a_0} \cdots$, and so

$$\frac{ax+b}{cx+d} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_r + x}}}, \quad r \text{ even}, \qquad \frac{ax+b}{cx+d} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_r + \cfrac{1}{x}}}}, \quad r \text{ odd}.$$

An immediate application is the remark about rational approximations of continued fractions. If $0 < x < 1$ is a real number, then finding the first n continued fraction digits of x computes $x = [0, a_1, \ldots, a_r + y]$, where $0 < y < 1$. This returns the fraction

$$\frac{p}{q} = [0, a_1, \ldots, a_r]\,.$$

By the above remarks (assume r is even) the form $[0, a_1, \ldots, a_r + y]$ corresponds to

$$\frac{p'y+p}{q'y+q}, \qquad \text{where } \det \begin{pmatrix} p' & p \\ q' & q \end{pmatrix} = 1\,.$$

So

$$x - \frac{p}{q} = \frac{p'y+p}{q'y+q} - \frac{p}{q} = \frac{y}{(q'y+q)q}\,.$$

In particular, the difference is $< 1/q^2$ which is a good approximation, since a rational number p/q will only be guaranteed to approximate within $1/q$. Furthermore, writing

$$y = \frac{1}{a_{r+1} + y'}$$

gives that the difference is less than

$$\frac{1}{a_{r+1}\, q^2}$$

which explains why cutting off the expansion before a large coefficient yields a good approximation.

This also allows one to prove a theorem due to Galois: *If $\alpha > 0$ has a purely periodic continued fraction expansion, then it is a root of a quadratic polynomial whose other root is $-1/\beta$, where $\beta > 0$ also has a purely periodic expansion whose period is the reverse of the period of α.*

Galois proved this when he was still in high school (interestingly, like Euclid, his proof consists of giving an example with 4 terms). The previous techniques now yield a simple proof: First, assume that $\alpha = [\overline{a_0, a_1, \ldots, a_r}]$, where the overline means it is periodic, and $a_0 > 0$, r odd. By the above, this implies that $A\alpha = \alpha$, where $A = U^{a_0} L^{a_1} \cdots L^{a_r}$. Now consider the other root $A\gamma = \gamma$, where $\gamma = -1/\beta$. Since $S\beta = -1/\beta$, it follows that $AS\beta = S\beta$, so that $S^{-1}AS\beta = \beta$. But a simple computation shows that

$$S^{-1}AS = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} a & c \\ b & d \end{pmatrix} = A^t \,,$$

so that $A^t\beta = \beta$. As before, A^t corresponds to the reverse continued fraction and thus $\beta = [\overline{a_r, \ldots, a_0}]$.

The result follows similarly if $a_0 = 0$ and r is even, but here one must define a purely periodic continued fraction to mean $[0, \overline{a_1, \ldots, a_r}]$. The other cases are left as an exercise.

One can use these same ideas to characterize the continued fraction expansion of $\sqrt{D}$, namely

$$\sqrt{D} = [\lfloor\sqrt{D}\rfloor, \overline{a_0, a_1, \ldots, a_1, a_0, 2\lfloor\sqrt{D}\rfloor}] \,,$$

where $a_0, a_1, \ldots, a_1, a_0$ is a palindrome. Note that this essentially says that the continued fraction is purely periodic with period $\lfloor\sqrt{D}\rfloor, a_0, \ldots, a_0, \lfloor\sqrt{D}\rfloor$, where the ends are straddling each other.

The solution of Pell's equation $a^2 - b^2 D = 1$ will be assumed, and this will illustrate the equivalence of the solution with b/d being a convergent to $\sqrt{D}$. Thus, let $\alpha = \sqrt{D}$ and $a^2 - b^2 D = 1$, where $a, b > 0$. If $A = \left(\begin{smallmatrix} a & bD \\ b & a \end{smallmatrix}\right)$, then A has determinant one, and $A\alpha = \alpha$. Since $bD > a$, $a > b$, one has the factorization $A = U^{a_0} L^{a_1} \cdots U^{a_r}$, where $a_r > 0$. This means that

$$\sqrt{D} = [a_0, a_1, \ldots, a_r + \sqrt{D}] \,.$$

As before, it follows that $A^t\beta = \beta$, where $\beta = 1/\sqrt{D}$, so that

$$\frac{1}{\sqrt{D}} = [0, a_r, a_{r-1}, \ldots, a_0 + 1/(1/\sqrt{D})] = [0, a_r, a_{r-1}, \ldots, a_0 + \sqrt{D}] \,,$$

and thus

$$\sqrt{D} = [a_0, a_1, \ldots, a_r + \sqrt{D}] = [a_r, a_{r-1}, \ldots, a_0 + \sqrt{D}] \,.$$

This shows that $a_0, \ldots, a_r$ is a palindrome, and the final form follows on noting that $a_0 = \lfloor\sqrt{D}\rfloor$.

7 Continuous versus discrete

Consider the following two questions:

C. How much information do n continued fraction digits of a number give you, i.e., what is the "base" corresponding to the continued fraction expansion?

D. How many division steps are there typically when the Euclidean algorithm is performed on two integers?

To analyze the first question, recall the basic setup of information theory: Given an unknown x, one measures the information content of an expansion by the amount one learns about x. Thus, if one knows n base 10 digits of x taken in $(0, 1)$, one has reduced x to an interval of size 10^{-n}. Knowledge of one more digit further reduces the interval to size 10^{-n-1}, therefore, one has narrowed x by a factor of 10. The Shannon McMillan Breiman theorem [5] then says that the *entropy* of the shift map $T_{10}(.a_1a_2\ldots) = .a_2a_3\ldots$ is $\log 10$.

In the case of continued fractions, one looks at an $x \in (0,1)$ and writes $x = [0, a_1(x), a_2(x), \ldots]$. The first n continued fraction digits give an approximation

$$\frac{p_n}{q_n} = [0, a_1(x), \ldots, a_n(x)]\,,$$

and, by the above,

$$\left| x - \frac{p_n}{q_n} \right| \le \frac{1}{q_n^2}\,.$$

If one accepts that typically this bound cannot be improved much, then knowledge of the first n digits narrows x to an interval of size about $1/q_n^2$. Knowledge of one more digit therefore gives a further refinement by a factor of $\frac{q_{n+1}^2}{q_n^2}$. Assuming that q_n grows exponentially, each continued fraction digit is worth about $\alpha = \lim_{n\to\infty} \frac{q_{n+1}^2}{q_n^2}$. If $\beta = \lim_{n\to\infty} \log q_n/n$ exists, then $\alpha = e^{2\beta}$, so question C is reduced to proving that β exists, i.e., finding the asymptotics of q from the length of the continued fraction.

Question D asks exactly the inverse questions as the number of steps in the Euclidean algorithm is just the length of the continued fraction expansion of p/q. In other words, one is looking at all continued fraction expansions $p/q = [0, a_1, \ldots, a_r]$, where $p < q$ and asking for r in terms of q.

It makes sense to believe that rational fractions are quite well behaved and that this question should reflect the behavior of real numbers. In other words, if for almost all real numbers, and expansion of length n has denominator of size $e^{n\beta}$, then for almost all p/q, one should have $r = (\log q)/\beta$. It will be seen that this is in fact the case.

The relationship between these two questions can be described by saying that question C is the continuous case and question D is the discrete case. This is completely analogous to the simpler example of base 10 digits. It is well known that for almost real numbers with base 10 expansion $x = 0.a_1a_2\ldots$, one has $a_1 + \cdots + a_n \sim 4.5\,n$, as $n \to \infty$ (continuous case). On the other hand, for almost all $n < N$, one has $a_0 + \cdots + a_r \sim 4.5 \log_{10} n$, where $n = \sum_{i=0}^{r} a_i 10^i$ (discrete case). In fact, both these results are special cases of the central limit theorem.

8 Statistics

To analyze the average growth of the q_n's one must first understand the general distribution of continued fraction digits, i.e., looking at a random number $0 < x < 1$, what is the distribution of the nth continued fraction digit? For example The distribution of the first digit is quite easy: The digit is k exactly when $k \le 1/x \le k+1$, i.e., $1/(k+1) \le x < 1/k$. To make things more precise, consider the *continued fraction map* $Tx = \{1/x\}$ which takes $[0, a_1(x), a_2(x), \ldots] \mapsto [0, a_2(x), a_3(x), \ldots]$, i.e., is the shift map for continued fraction expansions. Everything will be answered once the measure $F_n(x)$ of the set $\{z : T^n(z) \le x\}$ is understood. Assuming for the moment that $F_n(x)$ is given, then $F_{n+1}(x)$ can be expressed as

$$F_{n+1}(x) = \sum_{k=1}^{\infty} F_n\left(k \le \frac{1}{T^n x} \le k + x\right) = \sum_{k=1}^{\infty}\left(F_n\left(\frac{1}{k}\right) - F_n\left(\frac{1}{k+x}\right)\right)\,.$$

The hope is that the F_n's approach a limit F, which will have to have the property that

$$F(x) = \sum_{k=1}^{\infty} \left(F\left(\frac{1}{k}\right) - F_n\left(\frac{1}{k+x}\right) \right) .$$

It turns out that the function $\log(1+x)$ satisfies this linear relation, while the condition $F(1) = 1$ implies that $\log_2(1+x)$ is the candidate of choice.

If this convergence conjecture is correct, then the probability that the nth coefficient is ≥ 1 is

$$F(1/2) = \frac{\log 4 - \log 3}{\log 2} \approx .58496 .$$

This allows one to predict that the probability that digit k appears is

$$\frac{1}{\log 2} \int_{1/k+1}^{1/k} \frac{dx}{1+x} = \frac{1}{\log 2} \log\left(1 + \frac{1}{k(k+2)}\right) ,$$

so for large n, one has

a quotient of 1 about $\frac{\log 4/3}{\log 2} \approx 41.504$ percent of the time,

a quotient of 2 about $\frac{\log 9/8}{\log 2} \approx 16.992$ percent of the time,

a quotient of 3 about $\frac{\log 16/15}{\log 2} \approx 9.311$ percent of the time.

In fact, the study of F_n goes back to Gauss who considered the problem in 1800. His notebook includes the four place value of $F_2(1/2)$. Gauss further wrote: " *Tam complicatae evadunt, ut nulla spes superesse videatur,*" which means "They come out so complicated that no hope appears to be left." Twelve years later, Gauss wrote to Laplace saying: "I found by very simple reasoning that, for n infinite, $F_n(x) = \log(1+x)/\log 2$. But the efforts which I made since then in my inquiries assign $F_n(x) - \log(1+x)/\log 2$ for very large but not infinite values of n were fruitless." Gauss did not publish his results and his conjecture remained unproved until 1928 when Kuz'min showed that

$$F_n(x) = \frac{\log(1+x)}{\log 2} + O(e^{-A\sqrt{n}}) .$$

This was subsequently improved by Lévy who showed that the error is $O(e^{-An}$, for some $A > 0$. Wirsing subsequently showed that the error term is asymptotic to $(-\lambda)^n \Psi(x)$, where $\lambda = .30366\ldots$ and Ψ is analytic in the complex plane except for $(-\infty, -1)$. A complete solution to Gauss' problem was found by Babenko who showed that

$$F_n(x) = \frac{\log(1+x)}{\log 2} + \sum_{j=1}^{\infty} \lambda_j^n \Psi_j(x)$$

where $\lambda_j \to 0$ are real, $|\lambda_2| > |\lambda_3| \geq |\lambda_4| \geq \cdots$, and each Ψ_j is analytic in the above region. It is conjectured that the λ_j's are simple and alternate in sign. This has been verified computationally for the first 37 eigenvalues [13] [57].

Babenko considers the derivative of formula $(*)$, i.e., analyzes the operator

$$G(f(x)) = \sum_{k=1}^{\infty} \frac{1}{(k+x)^2} f\left(\frac{1}{k+x}\right) .$$

This has eigenfunction $1/(1+x)$ with eigenvalue 1. In general, the operator

$$G_s(f(x)) = \sum_{k=1}^{\infty} \frac{1}{(k+x)^{2s}}\, f\left(\frac{1}{k+x}\right)$$

is related to the spectral theory of $SL(2,\mathbf{Z})$ via the relation

$$Z(s) = \det(1-G_s)\ \det(1+G_s)\,,$$

where $Z(s)$ is the ordinary Selberg zeta function for the modular group. In particular, this shows that the operator G_s has has eigenvalue 1 exactly when $s = 1/2 + ir$ and $1/4 + r^2$ is an eigenvalue of the Laplacian on $\mathbf{H}/SL(2,\mathbf{Z})$, so that $s = 1$ corresponds to the eigenvalue $\rho = 0$, i.e., the constant functions.

Even though Gauss' problem has been completely solved, our question can be answered without knowing about the asymptotic properties of F_n. In fact, all one needs to know is that the *Gauss measure*

$$\frac{1}{\log 2}\int \frac{dx}{1+x}$$

is invariant under the map $x \mapsto Tx$ and that the easily proved fact that the only invariant sets have measure zero or one. This says that the continued fraction map is *ergodic* with respect to Gauss measure. The ergodic theorem then implies that for almost all x

$$\lim_{N\to\infty} \frac{f(Tx) + f(T^2x) + \cdots + f(T^Nx)}{N} = \frac{1}{\log 2}\int_0^1 f(x)\,\frac{dx}{1+x}\,,$$

in other words, averages along the first continued fraction digit over all x are reflected as averages along all digits for a single x. This is an example of the "ergodic hypothesis" which says that time averages reflect space averages. For example, one would expect that if a room is divided into two equal parts, then an air molecule should spend $1/2$ of its time in each part. Another example of the ergodic hypothesis is that if 10% of the population is over 6ft. tall, then a person should be over 6ft. tall for 10% of his life.

The asymptotic formula for $F_n(x)$ enables one to understand a lot about the statistical properties of continued fraction digits (note that since Gauss measure is absolutely continuous with respect to ordinary measure, "almost all" results apply to Lebesgue measure as well). If $f(x)$ is a function of the digits, then the ergodic theorem says that

$$\lim_{n\to\infty} \frac{f(a_1(x)) + \cdots + f(a_n(x))}{n} = \sum_{k=1}^{\infty} \frac{f(k)}{\log 2}\,\log\left(1+\frac{1}{k(k+2)}\right)\,.$$

This immediately proves Kinchin's result that the geometric mean of the digits tends to a limit

$$\sqrt[n]{a_1a_2\cdots a_n} \to \prod_{k=1}^{\infty}\left(1+\frac{1}{k(k+2)}\right)^{\log k/\log 2} = 2.68545200106530644530\ldots$$

One can similarly look at the harmonic mean defined by

$$H(v_1,\ldots,v_n) = \frac{n}{\dfrac{1}{v_1}+\cdots+\dfrac{1}{v_n}}\,.$$

The harmonic mean answers the question: "If you travel over n equal distances each at speed v_i, what is your average speed?" (The fact that $H(v_1, v_2) < (v_1 + v_2)/2$ when $v_1 \neq v_2$ is the reason why cycling out and back when there is wind is slower than when it is calm.) For the case of continued fraction coefficients, the harmonic mean on average is

$$\begin{aligned}\lim_{n\to\infty} \frac{n}{\frac{1}{a_1} + \cdots + \frac{1}{a_n}} &= \frac{\log 2}{\sum_{k=1}^{\infty} \frac{1}{k} \log\left(1 + \frac{1}{k(k+2)}\right)} \\ &= \frac{1}{\frac{3}{2} + \frac{2}{\log 2} \sum_{n=3}^{\infty} (2^{n-2} - 1)\left(\zeta'(n) + \frac{\log 2}{2^n}\right)} \\ &= 1.74540566224073468634\ldots\end{aligned}$$

where the evaluation in terms of values of $\zeta'(k)$ is done as in [81, Chapter 8]. Gosper [33] has computed the harmonic mean of the first million digits of π to be about 1.745942 and for the first 17 million digits about 1.745882. Since $e = [2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, \ldots]$ the harmonic mean of its digits is $3/2$.

9 Answers to C and D

The growth of $q_n(x)$, and therefore Question C, can now be analyzed using a very simple trick. Note that when the Euclidean algorithm is applied to p/q that the denominator of the last iteration become the numerator of the next iteration, so that multiplying the successive fractions gives

$$\frac{p_1}{q_1} \frac{p_2}{q_2} \cdots \frac{p_n}{q_n} = \frac{p_1}{q_n}\,.$$

On the other hand, $p_k(x)/q_k(x)$ is a very good approximation for $T_k x$, so that

$$\frac{\log q_n}{n} \approx -\frac{\log T_1 x) + \cdots + \log T_n x}{n} \to \frac{1}{\log 2} \int_0^1 (-\log x)\, \frac{dx}{1+x}\,.$$

The integral can be evaluated as follows. Consider the Γ function $\Gamma(s) = \int_0^\infty e^{-t} t^{s-1} dt$, then $\frac{\Gamma(s)}{n^s} = \int_0^\infty e^{-nt} t^{s-1} dt$, so that

$$\Gamma(s) \sum_{n=1}^{\infty} \frac{a_n}{n^s} = \int_0^\infty \left(\sum_{n=1}^{\infty} a_n e^{-nt}\right) t^{s-1} dt\,. = \int_0^1 \left(\sum_{n=1}^{\infty} a_n x^n\right) (-\log x)^{s-1} dx\,.$$

For $a_n = (-1)^{n+1}$ then the infinite series represents a rational function giving

$$\Gamma(s) \left(1 - \frac{1}{2^{s-1}}\right) \zeta(s) = \int_0^1 (-\log x)^{s-1} \frac{dx}{1+x}\,,$$

and in particular, this gives

$$\int_0^1 (-\log x)\, \frac{dx}{1+x} = \frac{1}{2}\, \zeta(2) = \frac{\pi^2}{12}\,.$$

The conclusion is that for almost all x

$$\beta = \lim_{n\to\infty} \frac{q_n}{n} = \frac{\pi^2}{12 \log 2}\,.$$

One can now answer Question C: Since each digit contributes

$$\frac{q_{n+1}^2}{q_n^2} \to e^{\pi^2/6\log 2} = 10.731\ldots,$$

times more information, continued fractions correspond to to expanding a number in base 10.731..., i.e., they are slightly more efficient than base 10 (though digits are not of fixed length, so one should really compare with Khinchin's constant). This number is therefore the entropy of the continued fraction map. It was first computed by Rohlin [67].

Question D can now also be answered. One must assume that the average behavior of integers $0 < p < q$ approximates that of random reals in $[0,1]$. As noted above, this would imply that $p/q = p_n/q_n$, where $q = q_n \approx e^{\pi^2 n/12\log 2}$ so that $\frac{12\log 2}{\pi^2}\log q$ should be the average running time of the Euclidean algorithm. In fact Heilbronn showed that this holds, and Knuth refined this to the formula

$$\frac{12\log 2}{\pi^2}\log N + C + O(N^{-1/6+\varepsilon}),$$

where

$$C = \frac{6\log 2}{\pi^2}(3\log 2 + 4\gamma - 24\pi^2\zeta'(2) - 1) - \frac{1}{2}.$$

This result is confirmed by Dixon's theorem which says that almost all $p < q < N$ satisfy

$$\left|\ell(p/q) - \frac{12\log 2}{\pi^2}\log q\right| < (\log q)^{1/2+\varepsilon}, \qquad \text{except for} \quad \exp(-c(\varepsilon)(\log N)^{\varepsilon/2})\,N^2 \text{ values},$$

where $\ell(p/q)$ is the length of the continued fraction of p/q. In fact, D. Hensley [41] has recently improved Dixon's result by showing that the error term is Gaussian, i.e., the number of steps in the Euclidean algorithm is asymptotically Gaussian with mean $\frac{12\log 2}{\pi^2}\log x$ and variance $C_1 \log x$, where

$$C_1 = \frac{1}{\pi^6}\left(\frac{12\log 2}{\pi^2}\right)^3 \frac{d^2}{dx^2}\log\lambda(s)\bigg|_{s=2},$$

where $\lambda(s)$ denotes a zeta function associated with continued fractions (see ?? below).

10 Sums of coefficients

The question of estimating $S_N = \sum_{n=1}^N a_n(x)$ is much harder since the ergodic theorem no longer applies:

$$\sum k\log\left(1 + \frac{1}{k(k+2)}\right) \approx \sum k\frac{1}{k^2} = \sum\frac{1}{k} = \infty.$$

However, since

$$\sum_{k=1}^N k\,\log_2\left(1 + \frac{1}{k(k+2)}\right) \sim \log_2 N$$

one might expect that $\frac{S_N}{N} \sim \log_2 N$ a.e., but this is false in general. In fact, a result of Borel and Bernstein states that if $\varphi(1), \varphi(2), \ldots$ is a sequence of positive integers then for almost all $x \in (0,1)$, $a_n(x) > \varphi(n)$ infinitely often if and only if $\sum 1/\varphi(n)$ diverges (see [5] for a proof). In particular, one has $a_n(x) > n\log n\log\log n$ infinitely often, for almost all x.

Another difficulty is that the $a_n(x)$ are *not* independent, and in fact this theory is one of the fundamental examples of sums of identically distributed non–independent random variables.

In spite of this, Khinchin did show [44] that for a fixed $\varepsilon > 0$

$$\lambda\left\{x \in (0,1) : \left|\frac{S_N}{N\log_2 N} - 1\right| \geq \varepsilon\right\} = O\left(\frac{1}{\varepsilon^2 \log N}\right), \tag{2}$$

where λ is Lebesgue measure. This result has been refined in two interesting ways. First, Diamond and Vaaler [20] correctly noted that the difficulty lies in the rare exceptionally large coefficients predicted by the Borel–Bernstein result, and that everything works well when the largest coefficient is removed. Thus, if one lets

$$S_N^* = \sum_{n<N} a_n - \max_{n<N} a_n ,$$

then $S_N^* \sim N\log_2 N$, a.e., where the error term can be given by $N(\log N)^{1-\delta}$, for some effective value $\delta > 0$ ($\delta = 1/16$ seems to work, see [84]).

Diamond and Vaaler also showed (see also [84]) that for $\alpha + \beta > 1$ then, with probability one, there are at most finitely many $a_p > N \log^\alpha N$ and $a_q > N \log^\beta N$ for distinct $p, q \leq N$. In other words, you cannot have two very large coefficients. Since the probability that $a_k > N\log N$ is about $1/(N\log N)$. Combining with Diamond and Vaaler's first result, this shows that (2) is optimal, i.e.,

$$\lambda\left\{x \in (0,1) : \left|\frac{S_N}{N\log_2 N} - 1\right| \geq \varepsilon\right\} = \Omega\left(\frac{1}{\log N}\right),$$

in other words, the left hand side is not $o()$ of the right hand side. One can even remove the absolute values on the left hand side, as Diamond and Vaaler's result shows that the error from asymptotic always comes from an extra large coefficient.

The other result is due to L. Heinrich [39] who analyzed the error term $S_N - N\log_2 N$ and showed that it was essentially linear, i.e., that $(S_N - N\log_2 N)/N$ has a limiting distribution. Thus, let $G(x)$ be the distribution function with with characteristic function

$$g(t) = \int_{-\infty}^{\infty} e^{itx} dG(x)$$

given by

$$g(t) = e^{-\frac{\pi}{2\log 2}|t|\left(1 + i\mathrm{sgn}(t)\frac{2}{\pi}\log|t|\right)},$$

then

$$\left|\mu\left\{\omega : \frac{1}{n}\sum_{k=1}^{n} a_k - \frac{\log n - \gamma}{\log 2} < x\right\} - G(x)\right| = O\left(\frac{(\log n)^2}{n}\right),$$

where $\gamma = .5772\ldots$ is Euler's constant and μ is the Gauss measure.

This type of result is a generalization of the central limit theorem developed by Khinchin and Lévy [46], [31], [42]. Given independent random variables $X_1, X_2, \ldots$ that converge to a distribution when normalized by $N^{1/\alpha}$, i.e., if

$$\left\{\frac{1}{N^{1/\alpha}}(X_1 + \cdots + X_N) < x\right\}$$

has a limiting distribution, then it is given by $G_{\alpha,\beta}(x,\lambda)$ whose characteristic function

$$g_{\alpha,\beta}(t,\lambda) = \int_{-\infty}^{\infty} e^{itx} dG_{\alpha,\beta}(x,\lambda)$$

satisfies

$$\log g_{\alpha,\beta}(t,\lambda) = -\lambda\,|t|^{\alpha}(1+i\beta\mathrm{sgn}(t)\,\omega(t,\alpha))\,,$$

where $0<\alpha\le 2$, $|\beta|\le 1$, $\ell>0$, and

$$\omega(t,\alpha)=\begin{cases}\frac{2}{\pi}\log|t|, & \alpha=1,\\ \tan(\alpha\pi/2) & \text{otherwise.}\end{cases}$$

For example, the case $\alpha=2$ is the usual normal distribution, while $\alpha=1$, $\beta=0$ is the so called Cauchy distribution. Only in the case $\alpha=2$ do these distributions have finite variance, and they have finite expectation only when $\alpha>1$.

If the continuous case models the discrete one, then Kinchin's result implies that the sum of continued fraction coefficients $S(p/q)$ should satisfy

$$S(p/q)\sim\frac{12}{\pi^2}\log q\log\log q\,,\quad \text{for almost all } p<q<N.$$

This has not yet been proved. In fact, there are some subtleties as Knuth and Yao [51] showed that the *average* value of $S(p/q)$ taken over all $0<p<q$ is

$$\frac{1}{q}\sum_{0<p<q}S(p/q)\sim\frac{6}{\pi^2}\,(\log q)^2\,.$$

What is going on is that a small number of large coefficients are inflating the average. The fact that most values of $S(p/q)$ are much smaller than the average was shown in [82]: For $1>\alpha>1/2$,

$$S(p/q)\le(\log q)^{1+\alpha},\quad p<q<N,\quad \text{with at most } O(N^2\,(\log N)^{1/2-\alpha+\varepsilon}) \text{ exceptions,} \tag{3}$$

where ε is a positive number that can be chosen arbitrarily but affects the constant in the $O(\ )$ term.

These results have implication to the running time of the non–Euclidean algorithm, which is essentially the sum of the odd continued fraction coefficients. According to the above, the average running time should be $\frac{3}{\pi^2}\,(\log q)^2$ while almost all fractions should have running time $\frac{6}{\pi^2}\log q\log\log q$. Thus, the non–Euclidean algorithm runs much longer than the usual Euclidean one, which is not immediately apparent from its definition.

11 Physics

The Kasner billiard is a discrete dynamical system that appears in models of the early universe, see [87], and in particular the "Mixmaster Universe", a chaotic model of the early universe.

One can define the dynamics as follows: One takes the unit circle $\mathbf{T}$ and surrounds it with an equilateral triangle with vertices at $\alpha_0=1+i\sqrt{3}$, $\alpha_1=-2$, and $\alpha_2=1-i\sqrt{3}$. The points of tangency are the three roots of unity $1,\omega,\omega^2$, where $\omega=\frac{-1+i\sqrt{3}}{2}$. The unit circle is thus divided into three open sets A_j, $j=0,1,2$, where $A_j=\{e^{2\pi i\theta}:\ \frac{j}{3}<\theta<\frac{j+1}{3}\}$. The map is defined as follows: A point P on the circle is mapped to a point $Q=K(P)$ by considering the closest corner α of the triangle to P, then drawing the line αP and letting Q be the other intersection. If P is tangent to the triangle, then $K(P)=P$, i.e., P is a fixed point.

The dynamics of this map are quite complicated, which is in part due to the fact that there are three parabolic fixed points $1,\omega,\omega^2$, in particular, it will be seen that the ergodic hypothesis is violated in a strong sense. A natural question to ask is whether the system spends an equal amount of time in A_0, A_1, A_2. It will be seen that this is both true (in space) and false (in time).

The dynamics of this map was studied by Andersson and Vardi [1]. The first point of this note is to show that the dynamical history of this map can be completely described by continued fraction expansions. In order to do this, one first makes the observation that the map T consists of inversions. For elementary properties of inversions, see [11].

In particular, consider three circles κ_j, $j = 0, 1, 2$, each of radius $\sqrt{3}$ and with k_j having center at α_j, the corresponding vertex of the triangle. Note that these circles are orthogonal to $\mathbf{T}$.

It is easy to see that the map K is made of inversions about the κ_j's. Thus, consider a point P belonging to A_j. Then inversion about κ_j preserves $\mathbf{T}$ since this circle is orthogonal to κ_j. Moreover, inversion also preserves the line $\alpha_j P$, since this line goes through the center of inversion. It follows that the inverse of P lies on $\mathbf{T}$ and on $\alpha_j P$, but is unequal to P. It follows that the inverse of P is equal to $K(P)$.

One can now simplify the picture significantly by performing an inversion with center 1. In particular, consider the map $f(z) = \frac{\sqrt{3}}{i}\frac{z+1}{z-1}$ (this is essentially the usual map identifying the Poincaré model of the hyperbolic plane with the upper half plane model). Then $f(1) = \infty$, $f(\omega) = -1$, $f(-1) = 0$, and $f(-\omega) = 1$. Moreover, f is essentially an inversion, so it maps circles and lines into circles and lines. It follows that $\mathbf{T}$ maps to the real axis, that κ_0 maps to the line $x = -1$, κ_1 maps to the unit circle $\mathbf{T}$, and κ_2 maps to the line $x = 1$.

For simplicity, one calls the resulting regions A, B, C. After inversion, it is seen that the Kasner map is

$$Tx = \begin{cases} 2 - x, & \text{if } x > 2, \\ -2 - x, & \text{if } x < -2, \\ 1/x, & \text{if } -1 \leq x \leq 1. \end{cases}$$

There is an obvious symmetry between positive and negative reals, so one can identify A and C and the positive and negative parts of B, so that the Kasner map now becomes

$$Tx = \begin{cases} x - 2, & \text{if } x > 2, \\ 2 - x, & \text{if } 2 > x > 1, \\ 1/x, & \text{if } 1 \leq x > 0. \end{cases}$$

This last formulation now gives an explicit description in terms of continued fractions.

Thus, a point $x > 0$ in B has a continued fraction expansion $x = [0, a_1, a_2, \ldots]$, and since $Tx = [a_1, a_2, \ldots]$, there is also a natural extension of this to A and C. In particular, the continued fraction expansion on A will be of the form $[a_0, a_1, \ldots]$, where $a_0 > 0$. The dynamics are as follows:

(i) If $a_0 > 0$ is even, then $[a_0, \ldots] \mapsto [a_1, \ldots]$ in $a_0/2$ steps consisting of A–C exchanges.

(ii) If $a_0 > 0$ is odd, then $[a_0, \ldots] \mapsto [a_2 + 1, \ldots]$ in $\lfloor a_0/2 \rfloor$ steps consisting of A–C exchanges and $2a_1$ A–B interchanges.

The case (i) is obvious. To prove (ii), note that if a_0 is odd, then $[a_0, \ldots] \mapsto [1, a_1, \ldots]$ in $\lfloor a_0/2 \rfloor$ A–C interchanges.

Now if $a_1 > 1$, then

$$T[1, a_1 + y] = 1 - [0, a_1 + y] = 1 - \frac{1}{a_1 + y} = \frac{a_1 - 1 + y}{a_1 + y} = \cfrac{1}{1 + \cfrac{1}{a_1 - 1 + y}} = [0, 1, a_1 - 1 + y],$$

so that $T^2[1, a_1, \ldots] = [1, a_1 - 1, \ldots]$. After $2(a_1 - 1)$ steps, one arrives at $[1, 1, a_2, \ldots]$. Now

$$1 - \cfrac{1}{1 + \cfrac{1}{a_2 + z}} = \frac{1}{a_2 + 1 + z},$$

so that $[1, 1, a_2, \ldots] \mapsto [a_2 + 1, \ldots]$ in two steps.

The distribution of large coefficients immediately shows that the dynamical system is not ergodic. In other words, a point does not spend an equal amount of time in A, B, C. In particular, it is clear that if t_A is the number of times the point is in A for time t, then t_A/t has as limit points all points in $[0, 1/2]$.

12 Alternating sums

It is surprising that the alternating sum of continued fraction coefficients plays an important role in a number of different settings. The reason is that the alternating sum corresponds to an additive character of $SL(2, \mathbf{Z})$. The alternating sum occurs in the theory of modular forms, as will be seen below, so it is natural that it also appear in geometry, in the study of the modular surface $SL(2, \mathbf{Z}) \backslash \mathbf{H}$. Thus, Guivarc'h and Le Jan [34] studied the winding number of geodesics on the modular surface, which comes down to the analysis of alternating sums of continued fraction coefficients. Alternating sums also occur implicitly in the work of Kirby and Melvin [47].

Thus, let $\hat{\gamma}_t$ be a geodesic on the modular surface, and $W_{\hat{\gamma}_t}$ the number of times the geodesic winds around the cusp. Then they show that under the normalized Liouville measure $\hat{\mu}$ on $\Gamma \backslash G$, $W_{\hat{\gamma}_t}/t$ converges to a Cauchy distribution as $t \to \infty$.

The relation to continued fractions is that this winding number, in the case of the modular group, is exactly the alternating sum of continued fraction coefficients. This is clear using the usual coding of geodesics [73]. In fact, Guivarc'h and Le Jan essentially prove that $(-a_1 + a_2 \cdots + (-1)^N a_N)/N$ converges to a Cauchy distribution with characteristic function $e^{-\pi|t|/(2\log 2)}$.

This result is predicted by Heinrich's theorem on the error term in the sum of coefficients, but Guivarc'h and Le Jan's proof also works for general congruence subgroups.

Thus, assuming that Heirich's result holds for even and for odd coefficients, one would have

$$\left\{ \frac{\sum_{2k \le N} a_{2k}}{N} < x \right\} \to G_{1,1}(x, \pi/(4\log 2)) - G_{1,1}(x, \pi/(4\log 2)),$$

$$\left\{ \frac{\sum_{2k+1 \le N} a_{2k+1}}{N} < x \right\} \to G_{1,1}(x, \pi/(4\log 2)) - G_{1,1}(x, \pi/(4\log 2)).$$

Now, the difference of two $G_{1,1}(x, \pi/(4\log 2))$'s is a $G_{1,0}(x, \pi/(2\log 2))$, see [92], and this is exactly a Cauchy distribution with parameter $\pi/(2\log 2)$, as proved by Guivarc'h and Le Jan.

12.1 The discrete case

Guivarc'h and Le Jan proved that in the continuous case, alternating sums have a limiting distribution. The discrete case can now be predicted by rescaling by the factor $\pi^2/(12\log 2)$. This would then give that

$\sum(-1)^i a_i$ has a Cauchy distribution with parameter $\frac{\pi}{2\log 2}\,\frac{12\log 2}{\pi^2} = \frac{6}{\pi}$, in other words

$$\lim_{N\to\infty} \frac{|\{0<d<c<N:\ \mathrm{GCD}\,(d,c)=1,\ \sum_i(-1)^i a_i\ < x\,\log c\}|}{|\{0<d<c<N:\ \mathrm{GCD}\,(d,c)=1\}|} = \frac{1}{\pi}\int_{-\infty}^{x} \frac{\frac{6}{\pi}}{\frac{6^2}{\pi^2}+y^2}\,dy\,, \tag{4}$$

and characteristic function $e^{-6|t|/\pi}$. As in the continuous case, such a result would follow from the discrete version of Heinrich's result

$$\lim_{N\to\infty} \frac{\left|\{0<d<c<N:\ \mathrm{GCD}\,(d,c)=1,\ \sum_i a_i - \frac{12}{\pi^2}\log c\log\log c < x\log c\}\right|}{|\{0<d<c<N:\ \mathrm{GCD}\,(d,c)=1\}|} = G_{1,1}\left(x, \frac{6}{\pi}\right)\,,$$

however this has yet to be proved (as noted above, even the main term $S(d/c) \sim \frac{12}{\pi^2}\log c\log\log c$ is still unproved).

A proof of (4) was found by I. Vardi in [82] but used methods that are completely independent of Heinrich or of Guivarc'h and Le Jan. This method is based on the theory of modular forms and the first step is to restate the problem in terms of Dedekind sums.

12.2 Dedekind sums

The Dedekind sum appears to be have been mistakenly defined and instead should have been defined to be the alternating sum of continued fraction coefficients. Historically, the Dedekind sum is defined for relatively prime integers d and c as

$$s(d,c) = \sum_{h=1}^{c-1} ((hd/c))\,((h/c))\,,$$

where the symbol $((x))$ signifies

$$((x)) = \begin{cases} 0, & \text{if } x \text{ is an integer,} \\ x - \lfloor x\rfloor - 1/2, & \text{otherwise.} \end{cases}$$

(See Figure 1.) This sum was discovered by Dedekind in 1876 [16] while editing Riemann's collected works [66]. He used this sum to express the functional equation of the Dedekind η function

$$\eta(z) = e^{\pi i z/12} \prod_{n=1}^{\infty}(1-e^{2\pi i n z})$$

which he proved satisfies

$$\log\eta\left(\frac{az+b}{cz+d}\right) = \begin{cases} \log\eta(z) + \frac{1}{2}\log(cz+d) + \frac{\pi i}{12}\,\Phi\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)\,, & \text{for } c>0, \\ \log\eta(z) + \frac{\pi i b}{12}\,, & c=0, \end{cases} \tag{5}$$

where $\mathrm{Im}\,(z)>0$, $g = \left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)$, a,b,c,d are integers satisfying $ad-bc=1$, and $\Phi\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)$ is a certain integer. Note that

$$\log\eta(z) = \frac{\pi i z}{12} - \sum_{m,n=1}^{\infty} \frac{e^{2\pi i m n z}}{m}\,,$$

so is holomorphic for $\mathrm{Im}\,z>0$.

In fact, Dedekind evaluated $\Phi\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)$ as

$$\Phi\begin{pmatrix} a & b \\ c & d\end{pmatrix} = -3 + \frac{a+d}{c} - 12s(d,c)\,, \tag{6}$$

with $s(d,c)$ as above (and $\Phi\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)=1$).

Using this functional equation Dedekind proved a fundamental identity for the Dedekind sum namely the reciprocity law

$$s(c,d)=\frac{c}{d}+\frac{d}{c}+\frac{1}{cd}-s(d,c)\,.$$

Since then numerous elementary proofs of this (not involving the theory of functions) have been given, see [64]. It should be noted that the Dedekind sum does have independent interpretations, e.g., Mordell [58] proved that it counts the number of lattice points in a tetrahedron.

Dean Hickerson proved the conjecture of Rademacher that the values of $(d/c, s(d,c))$ are dense in the plane (note that $s(d,c)$ can be considered to be a function of c/d since $s(c,d)=s(c/(d,c),d/(c,d))$, as is easily shown). He proved this by giving a continued fraction formula for the Dedekind sum.

Note that the definition of the Dedekind sum gives that

$$s(d,c)=s(d\bmod c,c)$$

and the reciprocity law relates the value of $s(d,c)$ to $s(c,d)$. It follows that $s(d,c)$ can be computed using the Euclidean algorithm so it should be expressible in terms of the continued fraction expansion of d/c. In fact, this is the statement of Hickerson's result.

Theorem 1. *Let $d<c$ and $0<a<c$ be such that $ad\equiv 1 \bmod c$, then if $[0,a_1,a_2,\ldots,a_r]$ is the regular continued fraction expansion of d/c with r odd then*

$$s(d,c)=\frac{1}{12}\left(-3+\frac{a+d}{c}-\sum_{i=1}^{r}(-1)^i a_i\right)\,.$$

(This corrects an error in [82].)

Exercise. Use Theorem 1 to prove the density result for Dedekind sums.

Theorem 1 shows that

$$s(d,c)=-\frac{1}{12}\sum_{i=1}^{r}(-1)^i a_i+O(1)\,,$$

so that a limiting distribution result for alternating sums is equivalent to a limiting distribution result for Dedekind sums modulo a factor of 12. The exact result is

$$\lim_{N\to\infty}\frac{|\{0<d<c<N:\ \mathrm{GCD}\,(d,c)=1,\ \ s(d,c)<x\,\log c\}|}{|\{0<d<c<N:\ \mathrm{GCD}\,(d,c)=1\}|}=\frac{1}{\pi}\int_{-\infty}^{x}\frac{\frac{1}{2\pi}}{\frac{1}{(2\pi)^2}+y^2}\,dy\,.$$

This is the form of the result which follows from the theory of modular forms, as will be seen below.

Proving Theorem 1 consists in showing that

$$\Phi(g_1g_2)=\Phi(g_1)+\Phi(g_2)$$

which is true when the c values of all three terms are non negative. Otherwise, one has

$$\Phi(g_1g_2)=\Phi(g_1)+\Phi(g_2)+6$$

as was proved by Rademacher. This immediately shows that $\Phi(U) = 1$, $\Phi(S) = -3$, and so $\Phi(L) = \Phi(USU) = -1$. Translating this into the matrix factorization explains the alternating sum of coefficients. Indeed, let $0 \le d < c$, $c > a, b \ge 0$ be such that $ad - bc = 1$, and let $d/c = [0, a_1, \ldots, a_r]$, where r is odd. By the above, one has

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix} = U^{a_1} L^{a_2} \cdots U^{a_r} ,$$

where $c/d = [a_1, \ldots, a_r]$. It follows that

$$\Phi \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \Phi(L^{a_r} U^{a_{r-1}} \cdots L^{a_1}) = a_r\,\Phi(L) + a_{r-1}\,\Phi(U) + \cdots + a_1 \Phi(L) = -a_1 + a_2 \cdots - a_r ,$$

as claimed. The idea of Rademacher's proof is that

$$\begin{aligned} \log \eta((g_1 g_2)(z)) &= r(z) + \Phi(g_1 g_2) \\ &= \log \eta(g_1(g_2(z))) = s(z) + \Phi(g_1) + \log \eta(g_2(z)) \\ &= s(z) + t(z) + \Phi(g_1) + \Phi(g_2) \end{aligned}$$

12.3 Limiting distributions

One says that an arithmetic function $f(n)$, $n = 1, 2, \ldots$, has a limiting distribution $F(x)$ if

$$\lim_{N \to \infty} \frac{1}{N} |\{n < N : \quad f(n) < x\}| = F(x) .$$

In other words, one takes a histogram of values of the function $f(n)$ and looks at its shape.

One method of showing that an arithmetic function has a limiting distribution is due to Paul Lévy [79]. Lévy's theorem says that if

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n<N} e^{itf(n)} = g(t) ,$$

and $g(t)$ is continuous at $t = 0$, then $f(n)$ has a limiting distribution $F(x)$, where

$$g(t) = \int_{-\infty}^{\infty} e^{itx} dF(x) ,$$

is the *characteristic function* of the distribution (this is simply the probabilist's terminology for the Fourier transform).

In order to prove the limiting distibution result for Dedekind sums (and thus for alternating sums of continued fraction coefficients) one applies the Lévy theorem to $s(d, c)/\log c$ which should give

$$\lim_{N \to \infty} \frac{\displaystyle\sum_{\substack{0<d<c<N \\ \mathrm{GCD}\,(d,c)=1}} e^{its(d,c)/\log c}}{|\{0 < d < c < N : \quad \mathrm{GCD}\,(d, c) = 1\}|} = e^{-|t|/(2\pi)} ,$$

where the right hand side corresponds to the characteristic functions of the Cauchy distribution

$$\int_{-\infty}^{\infty} \frac{\alpha\, e^{ity}}{\alpha^2 + y^2}\, dy = e^{-\alpha|t|} ,$$

with $\alpha = 1/(2\pi)$. The well known estimate [37]

$$|\{0 < d < c < N : \ \mathrm{GCD}\,(d,c) = 1\}| = \frac{3N^2}{\pi^2} + O(N \log N)\,,$$

shows that Lévy's criterion can be rewritten as

$$\sum_{\substack{0<d<c<N \\ \mathrm{GCD}\,(d,c)=1}} e^{its(d,c)/\log c} \sim e^{-|t|/(2\pi)}\, \frac{3N^2}{\pi^2}\,. \tag{7}$$

Proving such a formula presents a number of technical difficulties. For example, one would like to remove the absolute values on the right hand side, and the bothersome $1/\log c$ term in the exponential. The first point can be taken care of by noting that $s(c-d,c) = -s(d,c)$ implies that the left hand side is independent of the sign of t. The second point can be taken care of by noting that the log function does not vary very much, and that for most values of $c < N$, $\log c$ is almost equal to $\log N$. An estimate using the continued fraction formula for Dedekind sums and the subsequent upper bound (3) of $S(d/c) \le (\log N)^{3/2+\varepsilon}$ for almost all $d < c < N$, shows that

$$\sum_{\substack{0<d<c<N \\ \mathrm{GCD}\,(d,c)=1}} e^{its(d,c)/\log c} = \sum_{\substack{0<d<c<N \\ \mathrm{GCD}\,(d,c)=1}} e^{its(d,c)/\log N} + O(N^2 (\log N)^{-1/5+\varepsilon})\,,$$

see [82] for details. The problem is therefore reduced to showing that

$$\sum_{\substack{0<d<c<N \\ \mathrm{GCD}\,(d,c)=1}} e^{its(d,c)/\log N} \sim e^{-t/(2\pi)}\, \frac{3N^2}{\pi^2}\,, \quad t > 0\,. \tag{8}$$

To prove this result, one rewrites the left hand side as a sum of generalized Kloosterman sums, and these sums can be estimated using the theory of non–holomorphic modular forms.

12.4 Kloosterman sums

The Kloosterman sum is defined in the simplest case by

$$S(m,n,c) = \sum_{\substack{d<c \\ ad\equiv 1 \ (\mathrm{mod}\ c) \\ \mathrm{GCD}\,(d,c)=1}} e^{2\pi i(ma+nd)/c}\,.$$

It was introduced by Kloosterman in 1927 as a refinement to the Hardy–Littlewood circle method. The Kloosterman sum has about c terms, but one should expect a lot of cancellation since the values of $d \bmod c$ and $d^{-1} \bmod c$ should be independent, so that the numbers $e^{2\pi i(ma+nd)/c}$ should be "randomly" distributed on the unit circle. If this is true, then there would be maximum cancellation in the Kloosterman sum, implying that the size of sum is about the square root of the number of terms, i.e.,

$$S(m,n,c) = O(c^{1/2+\varepsilon})\,, \tag{9}$$

for any fixed $\varepsilon > 0$. Work of Salié [68] then Davenport [14] approached this estimate, which was finally proved by A. Weil in 1948 [89]. This proof essentially came down to Weil's proof of the Riemann hypothesis for curves.

Much recent progress in analytic number theory (the so-called "Kloostermania") was made by using estimates not just on Kloosterman sums, but on the fact that there is further cancellation when one sums Kloosterman sums [19]. These efforts have recently led to some spectacular results, in particular, the recent proof by J. Friedlander and H. Iwaniec that there are an infinite number of primes of the form $X^2 + Y^4$ [30, p. 952]: "We also do not appeal to the theory of automorphic functions although experts will, in several places, detect it bubbling just beneath the surface."

The basic observation is that there should be further cancellation if one looks at sums of Kloosterman sum. By the Weil estimate, one already has

$$\sum_{c<x} \frac{S(m,n,c)}{c} = O(x^{1/2+\varepsilon})\,.$$

However, it was proved by Kuznetsov [52] that the estimate holds when 1/2 is replaced by 1/6,

$$\sum_{c<x} \frac{S(m,n,c)}{c} = O(x^{1/6+\varepsilon})\,. \tag{10}$$

The proof depends strongly on modular forms and on the "Kuznetsov trace formula" described below.

In fact, Linnik [55] and Selberg [71] independently conjectured that the constant can be replaced by zero, i.e., that

$$\sum_{c<x} \frac{S(m,n,c)}{c} = O(x^{\varepsilon})\,,$$

for any $\varepsilon > 0$, but this has remained open.

The statement of the asymptotic formula for sums of Kloosterman sum will require the theory of modular forms.

12.5 Modular forms

The theory of modular forms (also called automorphic forms) can be thought of as a generalization of harmonic analysis to higher rank groups. In other words, classical harmonic analysis is the study of functions of the circle $\mathbf{R}/\mathbf{Z}$, while modular forms will study functions on a higher rank group modulo a discrete subgroup, for example, $SL(2,\mathbf{R})/SL(2,\mathbf{Z})$.

It is natural to begin with the classical theory of (holomorphic) modular forms, see [35] [69] [74]. We begin by giving the simplest examples of modular forms: holomorphic modular forms of even integral weight.

Let $\mathbf{H} = \{x+iy : \ y>0\}$ be the upper half plane. We will consider holomorphic functions on $\mathbf{H}$ that satisfy certain functional equations with respect to the action $z \mapsto \frac{az+b}{cz+d}$. The matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ will be restricted to certain finite index subgroups of $SL(2,\mathbf{Z})$ called *congruence subgroups*, defined as subgroups $G \subset SL(2,\mathbf{Z})$ containing $\Gamma(N)$ for some N, where

$$\Gamma(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbf{Z}) : \ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod N \right\}.$$

Note that $\Gamma(1) = SL(2,\mathbf{Z})$, so that G is a congruence subgroup if and only if $\Gamma(N) \subset G \subset \Gamma(1)$, for some N.

Modular forms satisfy a functional equation under the action of G, so their values are determined by a *fundamental domain* of $\mathbf{H}$ under the action of G which is denoted by $G\backslash\mathbf{H}$ (the upper half plane under this identification can also be thought of as a Riemann surface). In the case of the full modular group $\Gamma(1)$, this can be taken to be the well known region

$$\mathcal{F} = \{z \in \mathbf{H} : \ -1/2 \leq \mathrm{Re}\,(z) < 1/2, \ |z| \geq 1/2\}\,.$$

The part of $\mathcal{F}$ going off to infinity is called the *cusp* and this name becomes more clear visually if one uses the equivalent domain $-1/\mathcal{F}$ (see the figure).

Note that if G is a congruence subgroup, then the fundamental domain $G\backslash\mathbf{H}$ is simply a finite union of images of $\mathcal{F}$, and so there are a finite number of cusps and once again $G\backslash\mathbf{H}$ can be thought of as a Riemann surface.

Definition. A *modular form* of weight k for $SL(2,\mathbf{Z})$ is a holomorphic function on $\mathbf{H}$ satisfying

(i) $$f(gz) = (cz+d)^k\, f(z), \quad g \in SL(2,\mathbf{Z})\,,$$

(ii) $f(z)$ is bounded in the cusp of $SL(2,\mathbf{Z})\backslash\mathbf{H}$.

(iii) $f(z)$ is a *cusp form* if it approaches zero at the cusp.

More generally, a modular form for a congruence subgroup G will satisfy $f(gz) = (cz+d)^k\, f(z)$ for $g \in G$ and $f(z)$ is bounded at every cusp of $G\backslash\mathbf{H}$.

In the case of the modular group, the identity $f(z) = f(z+1)$ holds, since $\left(\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right)$ belongs to $SL(2,\mathbf{Z})$, so there is a Fourier development

$$f(x+iy) = \sum_{n=-\infty}^{\infty} a_n \psi_n(y) e^{2\pi i n x}\,.$$

The fact that $f(x+iy)$ is holomorphic implies that $\psi_n(y) = e^{-2\pi n y}$, and condition (ii) implies that $a_n = 0$ for $n < 0$. In other words, one has the Fourier expansion

$$f(z) = \sum_{n=0}^{\infty} a_n e^{2\pi i n z}\,. \tag{11}$$

Condition (iii) simply says that $a_0 = 0$.

The modular relation (i) above implies that modular forms of weight k correspond to meromorphic differentials of weight $k/2$ on $SL(2,\mathbf{Z})$, i.e., $f(z)(dz)^{k/2}$ is the set of all such differentials. The Riemann–Roch theorem then implies that the set of holomorphic modular forms of weight k is finite dimensional.

Examples. The Ramanujan–Δ function

$$\Delta(z) = e^{2\pi i z} \prod_{n=1}^{\infty} (1 - e^{2\pi i n z})^{24} = \sum_{n=1}^{\infty} \tau(n)\, e^{2\pi i n z}$$

is a modular form of weight 12. The fact that $\Delta(z)$ is a modular form is not trivial, see [69] for a proof, but it should be noted that $\Delta(z) = (\eta(z))^{24}$, so this already follows from the transformation law for $\eta(z)$ given above.

Modular forms also appear as a way of proving that a Dirichlet series

$$L_f(s) = \sum_{n=0}^{\infty} \frac{a_n}{n^s}$$

has a functional equation $s \mapsto k - s$. The idea is to show that the associated function

$$f(z) = \sum_{n=0}^{\infty} a_n \, e^{2\pi i n z} \,,$$

has a functional equation $f(-1/z) = z^k \, f(z)$. In fact, the functional equation $k \mapsto k - s$ then follows from the Mellin transform

$$\frac{\Gamma(s)}{(2\pi)^s} \, L_f(s) = \int_0^{\infty} [f(iy) - f(\infty)] y^s \, \frac{dy}{y} \,,$$

as was proved by Hecke. This idea goes back to Riemann who showed that the functional equation $s \mapsto 1 - s$ for $\zeta(s)$ follows from the functional equation for the theta function

$$\tilde{\theta}(z) = \sum_{n=-\infty}^{\infty} e^{\pi i n^2 z} \,,$$

given by $\tilde{\theta}(-1/z) = (-iz)^{1/2} \, \tilde{\theta}(z)$.

A general method of constructing modular forms is to average over the group G. The whole group is too large, since $f(z+1) = f(z)$, so one mods out such elements: Let

$$G_\infty = G \cap \left\{ \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} \in G \right\} \,,$$

then the Eisenstein series is defined to be

$$E_k(z) = \sum_{g \in G_\infty \backslash G} \frac{1}{j(g,z)^k} \,,$$

where $j(g,z) = cz + d$ when $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. The form of the sum shows that $E_k(z)$ is G invariant. In the case of $SL(2, \mathbf{Z})$, this is simply

$$E_k(z) = \sum_{\substack{c,d>0 \\ \mathrm{GCD}\,(c,d)=1}} \frac{1}{(cz+d)^k} \,,$$

and one can easily show that this sum converges to a holomorphic function in $\mathbf{H}$, see [69]. More generally, one has the Poincaré series

$$P_m(z,k) = \sum_{g \in G_\infty \backslash G} \frac{e^{2\pi i g z}}{j(g,z)^k} \,,$$

where $e^{2\pi i g z}$ is well defined modulo G_∞, and the sum converges as before.

Finally, the space of modular forms carries a natural inner product called the *Petersson inner product* defined by

$$\langle f, g \rangle = \iint_{G \backslash \mathbf{H}} f(z) \overline{g(z)} y^k \, \frac{dx dy}{y^2} \,.$$

Note that $\frac{dxdy}{y^2}$ is a measure which is invariant under the action of $SL(2, \mathbf{R})$. This follows from the identity

$$\mathrm{Im}\,(gz) = \frac{y}{|cz+d|^2} \tag{12}$$

which will be used repeatedly. A basic computation is

$$\langle P_m(\cdot,k), f\rangle = \iint\limits_{G\backslash\mathbf{H}} P_m(z,k)\overline{f(z)}y^k\,\frac{dxdy}{y^2}$$

$$= \iint\limits_{G\backslash\mathbf{H}} \sum_{g\in G_\infty\backslash G} j(g,z)^{-k}e^{2\pi i m g z}\overline{f(z)}y^k\frac{dxdy}{y^2}$$

One now observes that

$$\overline{f(z)} = \frac{\overline{f(gz)}}{\overline{j(g,z)^k}}\,,$$

so that

$$j(g,z)^{-k}\overline{f(z)}y^k = \frac{\overline{f(gz)}}{|cz+d|^{2k}}\,y^k = \overline{f(gz)}\,(\mathrm{Im}\,(gz))^k\,.$$

The above inner product is then

$$\iint\limits_{G\backslash\mathbf{H}} \sum_{g\in G_\infty\backslash G} e^{2\pi i m g z}\overline{f(gz)}(\mathrm{Im}\,(gz))^k\frac{dxdy}{y^2} = \iint\limits_{G_\infty\backslash\mathbf{H}} e^{2\pi i m z}\overline{f(z)}y^k\frac{dxdy}{y^2},$$

where the reduction follows from the fact that one is integrating over all copies of the fundamental domain modulo the transformation $z \mapsto z+1$ and the union of all these fundamental domains is therefore the fundamental domain for $G_\infty\backslash\mathbf{H}$. This this can be given by the region $0 \le x < 1$, $y > 0$, the integral equals

$$\int_0^\infty \int_0^1 e^{2\pi m z}\overline{f(z)}y^k\frac{dxdy}{y^2} = \frac{\overline{a_m}}{(4\pi m)^{k-1}}\,\Gamma(k-1)\,.$$

In other words, taking an inner product with $P_m(z,k)$ essentially gives the mth Fourier coefficient of $f(z)$. Since a function which has all zero Fourier coefficents is zero, this means that Poincaré series span the space of mdoular forms. Moreover, this shows that Eisenstein series correspond to the constant term as they are orthogonal to all cusp forms.

12.6 Success stories

Modular forms were the source of two of the most celebrated results of the 20th century: The Ramanujan–Petersson conjecture solved by Deligne in 1974 and the Shimura–Taniyama–Weil conjecture solved by Wiles, Breuil, Conrad, Diamond, and Taylor in 1999. As motivation for the subject, this section, which is independent of the rest of the paper, gives a brief description of these results.

The importance of modular forms comes in large part from the fact that their Fourier coefficients contain arithmetic information. Thus, Fourier coefficients of modular forms can be used to study the representation of integers as a sum of squares. The starting point is the θ function

$$\theta(z) = \sum_{n=-\infty}^{\infty} e^{2\pi i n^2 z}\,.$$

Since $\theta(z) = \tilde\theta(2z)$, the transformation law for $\tilde\theta(z)$ implies that

$$\theta(gz) = \left(\frac{c}{d}\right)\,\varepsilon_d^{-1}\,(cz+d)^{1/2}\theta(z)\,,\quad g\in\Gamma_0(4)\,,$$

and $\left(\frac{c}{d}\right) = \pm 1$ is a generalization of the Legendre symbol [69] and ε_d is 1 if $d \equiv 1 \pmod 4$ and i if $d \equiv 3 \pmod 4$. This implies that if k is divisible by 4, then $\theta^k(z)$ is a modular form for $\Gamma_0(4)$ of weight $k/2$. On the other hand,

$$\theta^k(z) = \sum_{n=0}^{\infty} r_k(n)\, e^{2\pi i n z}\,,$$

where $r_k(n)$ is the number of ways of writing n as a sum of k squares of integers.

As noted in the previous section, the space of modular forms of a given weight is finite and is spanned by Poincaré series. Moreover, the space of non–cusp forms is spanned by the Eisenstein series. In other words, every modular form can be written in the form $e + h$, where e is in the space spanned by Eisenstein series, and h is a cusp form. Now, it is simple to compute the Fourier series of Eisenstein series [35] [69] [74]. For example [74],

$$E_4(z) = 1 + 240 \sum_{n=1}^{\infty} \sigma_3(n) e^{2\pi i n z}\,, \qquad E_6(z) = 1 - 504 \sum_{n=1}^{\infty} \sigma_5(n) e^{2\pi i n z}\,,$$

where $d_k(n) = \sum_{d|n} d^k$. In general, the Fourier coefficients of Eisenstein series will be some type of finite divisor sum.

It follows that when the space of modular forms is 1–dimensional, then one gets a simple closed form for the number of representations of an integers as a sum of squares. This is in fact true for weight 2 and weight 4, and one has the well known formulas [37]

$$r_4(n) = 8 \sum_{\substack{d|n \\ 4\nmid n}} d\,, \qquad r_8(n) = (-1)^n\, 16 \sum_{d|n} (-1)^d d^3\,.$$

In general, such formulas will hold if there are no cusp forms. However, if cusp forms exist, then the divisor sums are now only main terms in asymptotics, while cusp forms correspond to error terms. From this point of view, it becomes important to understand the growth rate of Fourier coefficients of cusp forms, since these control the error term. This exact question was raised by Ramanujan who conjectured that for a cusp form of weight k, the Fourier coefficients of cusp forms should satisfy $a_n = O(n^{(k-1)/2+\varepsilon})$. Since the main term given by Eisenstein series has order n^{k-1}, this gives a square root error, which can be considered optimal. In fact, Ramanujan only made his conjecture for $\tau(n)$, and the general conjecture for cusp forms for congruence subgroups was made by Petersson, so this is now called the Ramanujan–Petersson conjecture. This conjecture was proved in Deligne's famous 1974 paper [17] in which he also proved the general Weil conjectures. Deligne's bound was later applied by Phillips and Sarnak to graph theory in the construction of explicit expanders [69]. For more information about representation of integers as sums of squares, see the survey article of W. Duke [22].

A more recent application of modular forms is Wiles' proof of Fermat's last theorem, [10] [90]. In particular, the proof followed from the proof by Taylor and Wiles of a special case of the Shimura–Taniyama–Weil conjecture. In fact, the proof of full conjecture has recently been completed by Breuil, Conrad, Diamond, and Taylor, see [12] for a very clear introduction, or [24] for more details.

The Shimura–Taniyama–Weil conjecture says that all rational elliptic curves are modular, which can be stated as follows: Given an elliptic curve E with rational coefficients, it is natural to construct a Dirichlet series

$$L(E,s) = \prod_{\substack{p \text{ prime} \\ p \nmid N}} (1 - a_p(E)p^{-s} + p^{1-2s})^{-1} = \sum_{n=0}^{\infty} \frac{a_n(E)}{n^s}\,,$$

where $N_p = p - a_p(E)$ is the number of points of E modulo p, except for a finite primes, those not dividing the *conductor* N. The conjecture is that there is a modular form of weight 2 for the congruence subgroup $\Gamma_0(N) = \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL(2,\mathbf{Z}) :\ c \equiv 0 \ (\mathrm{mod}\ N) \right\}$ given by

$$\sum_{n=1}^{\infty} a_n(M) e^{2\pi i n z}\,,$$

where $a_p(E) = a_p(M)$ for all primes $p \nmid N$.

Given the above results, it is interesting that, at first, there was some difficulty in accepting modular forms, as is described by Atle Selberg [72]:

> Littlewood and Hardy were primarily working with hard analysis and they did not have a strong feeling for modular forms and such things; the generating function for the partition is essentially a modular form, particularly if one puts the extra factor $x^{-1/24}$ to the power series. This must have been something that came quite naturally to Ramanujan from the beginning. But to Littlewood, in this review, it seems as if it was an afterthought by a particular stroke of genius that happened later in the development. I find this completely misleading...

Here, Selberg is referring to the partition function identity

$$\prod_{n=1}^{\infty} \frac{1}{1-x^n} = \sum_{n=1}^{\infty} p_n\, x^n\,,$$

where the left hand side is seen to be $1/\eta(z)$, except for the factor $x^{-1/24}$.

12.7 Non–holomorphic modular forms

Non–holomorphic modular forms were first introduced by Maass but the treatment which follows is due to A. Selberg [71]. As was alluded to above, there is an analogy with classical harmonic analysis in which one analyzes functions on the circle $\mathbf{R}/\mathbf{Z}$ using the Laplacian operator $\Delta = \frac{d^2}{dx^2}$, i.e., $\sin \lambda x$ and $\cos \lambda x$ will be eigenfunctions of Δ for a discrete set of λ.

Thus, one again looks at functions on the upper half plane which satisfy a transformation law under a congruence subgroup $G \subset SL(2,\mathbf{Z})$:

Definition. A *non-holomorphic modular form* of *weight* r and *multiplier system* χ is a function $f(z)$ on $\mathbf{H}$ satisfying

(i′) $$f(gz) = \chi(g) \left(\frac{cz+d}{|cz+d|} \right)^r f(z), \quad g \in G\,,$$

(ii′) $$\iint\limits_{G \backslash \mathbf{H}} |f(z)|^2 \frac{dx\,dy}{y^2} < \infty\,.$$

(iii′) $f(z)$ is a *cusp form* if it approaches zero at the cusp.

The first condition is simply a normalization of the functional equation of the previous section. In other words, if $F(z)$ satisfies $F(gz) = \chi(g)(cz+d)^r F(z)$, then the identity (12) implies that $f(z) = y^{k/2}F(z)$ satisfies condition (i$'$) above.

The second condition shows that these modular forms form a Hilbert space $L^2(G\backslash\mathbf{H},\chi,r)$ under the Petersson inner product

$$\langle f,g\rangle = \iint\limits_{G\backslash\mathbf{H}} |f(z)|^2 \frac{dxdy}{y^2}\,.$$

In order to understand this space, one decomposes it using the Laplacian operator

$$\Delta_r = y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right) - iry\,\frac{\partial}{\partial x}\,,$$

which is seen to preserve condition (i$'$), i.e., it is invariant under the action of G. One now looks at eigenfunctions and eigenvalues of this operator

$$\Delta_r u(z) + \lambda\, u(z) = 0\,.$$

Since Δ_r has a self adjoint extension to $L^2(G\backslash\mathbf{H},\chi,r)$, its spectrum is real and one has a sequence of eigenvalues going to infinity, with only a finite set of negative eigenvalues which correspond to holomorphic modular forms if r is an even integer. The non negative eigenvalues are simple except that the case $\lambda = 1/4$ could have multiplicity 2.

According to Selberg's notation, one writes an eigenvalue as $\lambda = s(1-s)$, with $\mathrm{Re}\,(s) \geq 1/2$. It follows that there a finite number of *exceptional* eigenvalues for which $\lambda < 1/4$. An exceptional eigenvalue corresponds to $s > 1/2$, while the other eigenvalues have $\mathrm{Re}\,(s) = 1/2$ (in analogy with the Riemann hypothesis).

Remark. Selberg conjectured that there are no exceptional eigenvalues for $\chi \equiv 1$ and weight zero for congruence subgroups (this can be considered a case of the Ramanujan–Petersson conjecture [69]). Selberg [71] showed that $\lambda \geq 3/16$ for congruence subgroups and this was recently improved by Luo, Rudnick, and Sarnak [56] to $\lambda \geq 21/100$.

For each exceptional eigenvalue λ_j, one considers the corresponding eigenfunction $u_j(z)$ normalized so that $\langle u_j,u_j\rangle = 1$. As with holomorphic modular forms, the transformation law yields a Fourier development, though the x and y terms no longer coincide to give an exponential. Thus, one uses separation of variables to get

$$u_j(z) = \rho_j(0)^{1-s_j}e^{-2\pi i\alpha x} + \sum_{n\neq 0}\rho_j(n)W_{(r/2)\mathrm{sign}\,(n-\alpha)/q,s_j-1/2}(4\pi|n-\alpha|y/q)e^{2\pi i(n-\alpha)x/q}\,,$$

where $q = \min\left\{n>0:\ \left(\begin{smallmatrix}1&n\\0&1\end{smallmatrix}\right)\in G\right\}$, $\chi\left(\begin{smallmatrix}1&q\\0&1\end{smallmatrix}\right) = e^{-2\pi i\alpha}$ with $0\leq\alpha<1$, and the $\rho_j(n)$ are the *Fourier coefficients* of the modular form ($\rho_j(0) = 0$, except possibly when $\alpha = 0$). Note that $W_{\beta,\mu}(y)$ is the Whittaker function which decays exponentialy in y as $y\to\infty$ and satisfies the O.D.E.

$$\frac{d^2W}{dy^2} + \left(-\frac{1}{4}+\frac{\beta}{y}+\frac{\frac{1}{4}-\mu^2}{y^2}\right)W = 0\,.$$

12.8 Generalized Kloosterman sums

Using the notation of the previous section, one can now define the generalized Kloosterman sum

$$S(m,n,c,\chi,G) = \sum_{\substack{0<a<qc \\ 0<d<qc \\ g=\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)\in G}} \overline{\chi(g)}\, e^{2\pi i((m-\alpha)a+(n-\alpha)d)/(qc)} ,$$

and give the estimate for sums of Kloosterman sums

$$\sum_{c<N} \frac{S(m,n,c,\chi,G)}{c} = \sum_{1/2<s_j<1} \tau_j(m,n,\chi,G)\, \frac{N^{2s_j-1}}{2s_j-1} + O(N^{\beta/3+\varepsilon}) , \tag{13}$$

where β is the best constant that one can put in the estimate

$$S(m,n,c,\chi,G) = O(c^{\beta+\varepsilon}) ,$$

and the sum is over exceptional eigenvalues, with

$$\tau_j(m,n,\chi,G) = \frac{q^2\,\overline{\rho_j(m)}\rho_j(n)(\pi^2(m-\alpha)(n-\alpha)/q^2)^{1-s_j}\Gamma(s_j+r/2)\Gamma(2s_j-1)}{(-i)^r\,\pi\Gamma(s_j-r/2)} . \tag{14}$$

This formulation of the Kuznetsov trace formula is due to Goldfeld and Sarnak [32]. In particular, one can read off the Kuznetsov estimate for sums of ordinary Kloosterman sums as follows. In the case of $SL(2,\mathbf{R})$, there are no exceptional eigenvalues, as was proved by Selberg (see [69] for a simple proof) so the sum is empty. Moreover, the Weil bound (9) says that one can put $\beta = 1/2$, which gives the result.

In the application to Dedekind sums, one will use almost every aspect of the definition of generalized Kloosterman sums, except that $G = SL(2,\mathbf{R})$ so that $q = 1$ in this case. In particular, the application will actually use the fact that there will be an exceptional eigenfunction, so that an asymptotic formula for sums of Kloosterman sums will be obtained. This appears to be the only case in which an asymptotic formula for such sums is required (as opposed to a simple upper bound).

The formula of Goldfeld and Sarnak is proved following ideas of Selberg. The heart of the proof is that the series

$$Z_{m,n}(s) = \sum_{c>0} \frac{S(m,n,c,\chi,G)}{c^{2s}}$$

determines the analytic character of the inner product of two generalized Poincaré series. The non–holomorphic Poincaré series is defined by analogy with the holomorphic case as

$$P_m(z,s,\chi,G) = \sum_{g\in G_\infty\backslash G} \frac{e^{2\pi i(m-\alpha)gz/q}}{j(g,z,\chi)}\, (\operatorname{Im} gz)^s , \tag{15}$$

where $j(g,z,\chi) = \chi(g)((cz+d)/|cz+d|)^r$, so $P_m(z,x,\chi,G)$ satisfies the transformation law (i′). If $m>0$, then $P_m(z,s,\chi,G)$ is square integrable, and one has a similar formula for its inner product

$$\langle P_m(\cdot,s,c,\chi,G),u_j\rangle = q\,\overline{\rho_j(m)} \left(\frac{4\pi(m-\alpha)}{q}\right)^{1-s} \frac{\Gamma(s+s_j-1)\Gamma(s-s_j)}{\Gamma(s-r/2)} .$$

Thus, taking an inner product with a Poincaré series picks out the mth Fourier coefficient of a non–holomorphic modular form.

The other important point is that Poincaré series also satisfy an identity with respect to the Laplacian

$$\Delta_r P_m(z,s,\chi,G) + s(1-s)P_m(z,s,\chi,G) = -4\pi\,\frac{m-\alpha}{q}\,P_m(z,s+1,\chi,G)\,,$$

as can be seen by applying the (invariant) Laplacian to every term in the sum (15). From this, Selberg concluded that $P_m(z,s,\chi,G)$ has a meromorphic continuation to $\mathrm{Re}\,(s) > 1/2$ with at most a finite number of simple poles at the points $s = s_j$, $1/2 < s_j < 1$, corresponding to the exceptional eigenvalues.

One now takes the inner product of two Poincaré series P_m and P_n, and this will be, in some sense, "universal" for pairs of Fourier coefficients. A computation [32] yields

$$\langle P_m(\cdot,s,\chi,G), P_n(\cdot,\bar{s}+2,\chi,G)\rangle = (-i)^r 4^{-s-1}\pi^{-1}\left(\frac{n-\alpha}{q}\right)^{-2}\frac{\Gamma(2s+1)}{\Gamma(s+r/2)\Gamma(s-r/2+2)}\,Z_{m,n}(s) + R(s)\,,$$

where $R(s)$ is holomorphic for $\mathrm{Re}\,(s) > 1/2$. Since the terms on the right hand side are meromorphic for $\mathrm{Re}\,(s) > 1/2$, it follows that $Z_{m,n}(s)$ will also have poles at the s_j's. Moreover, rates of growth for $Z_{m,n}(\sigma+it)$ can be estimated [32], and the residues at the poles s_j can be evaluated, so standard methods of number theory, e.g., as in the proof of the prime number theorem [15], can be applied to give the asymptotic formula for partial sums of the coefficients of $Z_{m,n}(s)$, i.e., equation (13).

12.9 Outline of proof

In order to prove our result, we look at the special case when $G = SL(2,\mathbf{Z})$, the weight r is a positive real number < 1, and χ_r is the character corresponding to the transformation law of the Dedekind–η function

$$\chi_r(g) = e^{2\pi i r\Phi(g)}\,,$$

with $\Phi(g)$ defined as in (6). The transformation law of the logarithm of $\eta(z)$ given by (5) shows that $y^{r/2}\eta^{2r}(z)$ is a non–holomorphic modular form of weight r for $SL(2,\mathbf{Z})$ with multiplier system χ_r and with $q=1$, $\alpha_r = \{-r/12\}$. The relationship with Kloosterman sums comes from a simple identity proved in [80]

$$e^{\pi i r/2}\sum_{\substack{0<d<c\\ \mathrm{GCD}\,(d,c)=1}} e^{2\pi i r s(d,c)} = S(1,1,c,\chi_r,SL(2,\mathbf{Z}))\,,$$

when $0<r<1$. One can therefore rewrite

$$\sum_{\substack{0<d<c<N\\ \mathrm{GCD}\,(d,c)=1}} e^{2\pi i r s(d,c)} = e^{-i\pi r/2}\sum_{0<c<N} S(1,1,c,\chi_r,SL(2,\mathbf{Z}))\,.$$

The right hand side is a sum of Kloosterman sums which can be estimated using the formula Goldfeld and Sarnak (one uses partial summation to remove the $1/c$ term in (13) changing the denominator on the RHS of (13) from $2s_j-1$ to $2s_j$). In order to do this, one must know all the exceptional eigenvalues for weight r and multiplier systems χ_r for $SL(2,\mathbf{Z})$. In fact, this was already done by R. Bruggeman [7] who showed that for $0<r<1$, $(r/2)(1-r/2)$ is the only exceptional eigenvalue. Moreover, a simple computation shows that this eigenvalue corresponds to $y^{r/2}\eta^{2r}(z)$. In order to make an explicit computation of the τ_j, one takes the normalized eigenfunction

$$u(z) = \frac{y^{r/2}\eta^{2r}(z)}{\sqrt{A_r}}\,,\quad \text{where } A_r = \iint\limits_{SL(2,\mathbf{Z})\backslash\mathbf{H}} y^r|\eta(z)|^{4r}\frac{dxdy}{y^2}\,.$$

It follows [82] that

$$u(z) = \sum_{n=1}^{\infty} \rho(n)\, (4\pi(n-\alpha_r))^{r/2}\, y^{r/2} e^{2\pi i(n-\alpha_r)x}\,,$$

where $\rho(1) = 1/(\sqrt{A_1}\,(4\pi(1-\alpha_r))^{r/2})$.

Substituting these values into (13) with $\beta = 1$ (the trivial bound) yields

$$\sum_{\substack{0<d<c<N \\ \mathrm{GCD}\,(d,c)=1}} e^{2\pi i r s(d,c)} = \frac{(1/4)^r}{\pi A_r(1-r/2)} N^{2-r} + O(N^{4/3+\varepsilon})\,.$$

To get the final result (8), one now puts $r = t/(2\pi \log N)$. This substitution gives

$$\sum_{\substack{0<d<c<N \\ \mathrm{GCD}\,(d,c)=1}} e^{its(d,c)/\log N} = \frac{\pi}{3}\frac{1}{A_0} e^{-t/(2\pi)} \frac{3N^2}{\pi^2} + O\left(\frac{1}{\log N}\right)\,.$$

The result now follows by the computation

$$A_0 = \iint_{SL(2,\mathbf{Z})\backslash \mathbf{H}} \frac{dxdy}{y^2} = \frac{\pi}{3}\,.$$

Remark. In the proof, the weight r always tends to zero, as t is fixed and $r = t/(2\pi \log N) \to 0$. It is also interesting to note that $1/(2\pi)$ appears naturally in the limiting distribution of Dedekind sums, as it represents the discrepancy between the 2π appearing from the multiplier system $\chi_r(g) = e^{2\pi i r g}$ arising from modular forms (harmonic analysis) and the exponential sum arising from Lévy's formula.

13 Arithmetic

Consider the problem of adding and multiplying the sequence of digits of two continued fraction expansions. Even multiplying a continued fraction by 2 is already a time consuming task based on "Hurwitz rules" [50, Exercise 4.5.3.14], so the general problem was declared to be hopeless by Khinchin, the renowned expert in the field [45], who wrote:

> There is, however, another and yet more significant practical demand that the aparatus of continued fractions does not satisfy at all. Knowing the representations of several numbers, we would like to be able, with relative ease, to find the representations of the simpler functions of these numbers (especially, their sum and product). In brief, for an apparatus to be suitable from a practical standpoint, it must admit sufficiently simple rules for arithmetical operations; otherwise, it cannot serve as a tool for calculation. We know how convenient systematic fractions are in this respect. On the other hand, for continued fractions there are no practically applicable rules for arithmetical operations; even the problem of finding the continued fraction for a sum from the continued fraction representing the addends is exceedingly complicated, and unworkable in computational practice.

The problem was solved in papers of Marshall Hall [36], G.N. Raney [65], and R.W. Gosper [3] [33] [50, Exercise 4.5.3.15] [53, p. 78]. The following is a description of Gosper's method, which is based on three ideas. The first idea is that a computation like $3x$, where x is given as a continued fraction, will rapidly lead to more complicated forms. For example, computing $3 \cdot \frac{14}{11}$

$$3\,[1,3,1,2] = 3\,\left(1 + \frac{1}{[3,1,2]}\right) = \frac{3\,[3,1,2]+1}{[3,1,2]}$$

and so forth. Gosper's idea was not to try to simplify these forms, but to analyze the worst that can happen. It's not hard to see that you always get a linear fractional form

$$\frac{ax+b}{cx+d}\,, \tag{16}$$

where a, b, c, d are integers, and x is given as a continued fraction.

Gosper's second idea was to consider the x as a *formal* symbol which could input continued fraction digits into (16). In other words, think of x not as representing a rational number, but as a symbolic quantity that transforms as

$$x \mapsto q + \frac{1}{x}\,, \tag{17}$$

where q represents the next continued fraction digit. It remains to see how (16) transforms under (17). A simple computation shows that

$$\frac{ax+b}{cx+d} \mapsto \frac{aqx+bx+a}{cqx+dx+c}\,.$$

So if the form (16) is represented by a matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, then the transformation law corresponding to inputting a digit is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} aq+b & a \\ cq+d & c \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} q & 1 \\ 1 & 0 \end{pmatrix}\,.$$

The third idea is to realize that you can output continued fraction coefficients without having complete knowledge of x. Gosper's original example was

$$\frac{70x+29}{12x+5}$$

If $x > 0$, then

$$\frac{29}{5} \le \frac{70x+29}{12x+5} < \frac{70}{12}\,.$$

This means that

$$\frac{70x+29}{12x+5} = 5 + \frac{10x+4}{12x+5}\,.$$

Similarly,

$$\frac{12}{10} \le \frac{12x+5}{10x+4} < \frac{5}{4}$$

so

$$\frac{12x+5}{10x+4} = 1 + \frac{2x+1}{10x+4}\,,$$

and

$$\frac{10x+4}{2x+1} = 4 + \frac{2x}{2x+1}\,.$$

This means that for *any* $x > 0$ you get a continued fraction expansion

$$\frac{70x+29}{12x+5} = 5 + \cfrac{1}{1 + \cfrac{1}{4 + \cfrac{1}{(2x+1)/x}}}$$

In other words, you are able to output 3 continued fraction coefficients of the form $(10x+25)/(2x+1)$ without knowing too much about x. (Note that you cannot output a fourth coefficient since $(2x+1)/x \mapsto 2+1/x$ does not give the correct coefficient unless $x > 1$.)

In the case when x represents a sequence of continued fraction coefficients, the situation is even better because the coefficients are greater or equal to one. Now, since one is assuming that there is a continuing sequence of continued fraction coefficients, the further assumption that $x > 1$ can be made, so being able to output a continued fraction corresponds to checking that for some integer n

$$n \leq \frac{a}{c} \leq \frac{a+b}{c+d} < n+1\,,$$

in other words, that

$$\left\lfloor \frac{a}{c} \right\rfloor = \left\lfloor \frac{a+b}{c+d} \right\rfloor\,,$$

and when this happens the output will be the common value $q = \lfloor a/c \rfloor$. The point is that is always happens, given a sufficient number of digits of x.

It remains to see what happens to the form after this has been output. This is

$$\frac{1}{\dfrac{ax+b}{cx+d} - q} = \frac{cx+d}{ax - cqx + b - dq}\,,$$

and the corresponding matrix transformation is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} c & d \\ a - cq & b - dq \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix}\,.$$

All these steps can be combined into an algorithm but first an algorithm to compute the continued fraction of a rational number must be written. This essentially follows the $\frac{14}{11}$ example (ordinary concatenation $[0] \star [x_0, \ldots, x_r] = [0, x_0, \ldots, x_r]$ will be used from now on).

Algorithm to compute the continued fraction expansion of a rational number: Given a rational number q, return a sequence $f(q)$:

if $q = 1/0$ **then** $[\,]$ **else** $[\lfloor q \rfloor] \star f(1/(q - \lfloor q \rfloor))$

Algorithm to compute the continued fraction expansion of $(ax+b)/(cx+d)$**:** Given a matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ and the continued fraction expansion $x = [x_0, \ldots, x_r]$, return the sequence $f(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), x)$:

if $r = 0$ **then** $f((a\,x_0 + b)/(c\,x_0 + d))$.

else if $a, b, c, d > 0$ and $\left\lfloor \frac{a}{c} \right\rfloor = \left\lfloor \frac{a+b}{c+d} \right\rfloor$ **then** $[\lfloor a/c \rfloor] \star f\left(\left(\begin{smallmatrix} 0 & 1 \\ 1 & -\lfloor a/c \rfloor \end{smallmatrix}\right) \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), x \right)$ (Output a digit.)

else $f\left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \left(\begin{smallmatrix} x_0 & 1 \\ 1 & 0 \end{smallmatrix}\right), [x_1, \ldots, x_r] \right)$ (Input a digit.)

Finally, the algorithms for addition and multiplication can be described. Just as before computing $x + y$ or xy leads to more complicated expressions. The worst expression you get is a "bilinear fractional form"

$$\frac{axy + bx + cy + d}{exy + fx + gy + h}\,. \tag{18}$$

Treating x and y as formal variables, this can be represented by an ordered pair of matrices (or tensor)

$$\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix}\right) .$$

Letting $y \mapsto q + 1/y$ in (18) gives a transformation for the tensor

$$\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix}\right) \mapsto \left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix}\right) \begin{pmatrix} q & 1 \\ 1 & 0 \end{pmatrix} ,$$

where ordinary matrix multiplication is used on each component. Similarly, letting $x \mapsto q + 1/x$ in (18) corresponds to

$$\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix}\right) \mapsto \begin{pmatrix} q & 1 \\ 1 & 0 \end{pmatrix} \left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix}\right) .$$

Just as before, one uses an Euclidean algorithm to output coefficients, but this time there won't be a guarantee that intermediate values will be greater than one. A tricky part of the algorithm is to decide whether to choose x or y to input the next digit. A direct solution to this is to alternate between them and this requires a componentwise transpose

$$\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix}\right)^T = \left(\begin{pmatrix} a & c \\ b & d \end{pmatrix}, \begin{pmatrix} e & g \\ f & h \end{pmatrix}\right) ,$$

corresponding to switching x and y in (18).

Algorithm to compute the continued fraction of $(axy + bx + cy + d)/(exy + fx + gy + h)$**:** Given as input a tensor $\left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), \left(\begin{smallmatrix} e & f \\ g & h \end{smallmatrix}\right)\right)$ and continued fraction expansions $x = [x_0, \ldots, x_r]$, $y = [y_0, \ldots, y_s]$, return a sequence $f\left(\left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), \left(\begin{smallmatrix} e & f \\ g & h \end{smallmatrix}\right)\right), x, y\right)$:

if $s = 0$ **then** $f\left(\left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), \left(\begin{smallmatrix} e & f \\ g & h \end{smallmatrix}\right)\right) \left(\begin{smallmatrix} y_0 \\ 1 \end{smallmatrix}\right), x\right)$.

else if $a, b, c, d, e, f, g, h > 0$ and $\lfloor a/e \rfloor = \lfloor b/f \rfloor = \lfloor c/g \rfloor = \lfloor d/h \rfloor$ **then**

$$[\lfloor a/e \rfloor] \star f\left(\left(\begin{smallmatrix} 0 & 1 \\ 1 & -\lfloor a/e \rfloor \end{smallmatrix}\right) \left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), \left(\begin{smallmatrix} e & f \\ g & h \end{smallmatrix}\right)\right), x, y\right) \qquad \text{(Output a digit.)}$$

$$\textbf{else} \;\; f\left(\left\{\left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), \left(\begin{smallmatrix} e & f \\ g & h \end{smallmatrix}\right)\right) \left(\begin{smallmatrix} y_0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\right\}^T, [y_1, \ldots, y_s], x\right) \qquad \text{(Input a digit and switch } x, y.\text{)}$$

This algorithm allows one to add and multiply continued fractions according to the rules

$$x \oplus y = f\left(\left(\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)\right), x, y\right) , \qquad x \otimes y = f\left(\left(\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)\right), x, y\right) .$$

Remark 1. The real advantage of Gosper's method is that it allows you to add and multiply numbers with known continued fraction expansions, for example, $e = [2, 1, 2, 1, 1, 4, 1, 1, 6, \ldots]$ and $\phi = \frac{\sqrt{5}+1}{2} = [1, 1, 1, \ldots]$. The above algorithms are easily modified to compute quantities like $e + \phi$ where the digit input would be the functions

$$e(n) = \begin{cases} 2 & \text{if } n = 0 \\ 2(n+1)/3 & \text{if } n \equiv 2 \pmod 3 \\ 1 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad \phi(n) = 1 .$$

Remark 2. It would be interesting to find a geometric interpretation of Gosper's method, as in [73], and also for bilinear fractional forms.

14 Analysis infinitorum

An entire series f can be expanded into continued fractions in various ways. The most important perhaps,

$$\sum_{n-=0}^{\infty} f_n z^n = \cfrac{c_0}{1 + b_1 z + \cfrac{c_1 z^2}{1 + b_2 z + \cfrac{c_2 z^2}{\ddots}}} \tag{19}$$

is called a Jacobi fraction or J-fraction. The case where all b_j are zero (together with related normalizations) gives rise to Stieltjes fractions or S-fractions; see [40, 59, 88]. Such expansions were called "algebraic continued fractions" in the 19th century; they are of considerable historical importance since they are at the origin of the theory of orthogonal polynomials [6].

Given a series f, a J-fraction expansion is obtained by an algorithm of the Euclidean type: define the integer and fractional parts of f by

$$\lfloor f \rfloor = f_0 + f_1 z, \qquad f = \lfloor f \rfloor + z^2 \{f\},$$

then iterate the transformation $f \mapsto \{1/f\}$. Accordingly, many continued fraction expansions of classical special functions derive from the iteration of functional relations involving quotients. Amongst these, the continued fraction of Gauss stands out. Gauss starts from the hypergeometric function classically defined by

$$F(\alpha, \beta, \gamma; z) = 1 + \frac{\alpha\beta}{\gamma}\frac{z}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{z^2}{2!} + \cdots$$

and derives by simple algebra

$$\frac{F(\alpha, \beta+1, \gamma+1; z)}{F(\alpha, \beta, \gamma; z)} == \cfrac{1}{1 - z\dfrac{\alpha(\gamma-\beta)}{\gamma(\gamma+1)}\dfrac{F(\alpha+1, \beta+1, \gamma+2; z)}{F(\alpha, \beta+1, \gamma+1; z)}}.$$

When iterated, this last functional relation expresses the quotient of two contiguous hypergeometric functions as a continued fraction, with the coefficients at level k that involve k and the parameters rationally.

Now, the hypergeometric function specializes in various ways to classical functions like $\log(1+z)$, $\arcsin z$, e^z, $J_0(z)$ (a Bessel function), etc. For instance, Gauss' fraction yields

$$e^z = \cfrac{1}{1 - \cfrac{z}{1 + \cfrac{z}{2 - \cfrac{z}{3 + \cfrac{z}{\ddots}}}}},$$

providing an arithmetical continued fraction for $e = \exp(1)$ from which the irrationality of e is apparent. The related expansion

$$\frac{\tan z}{z} = \cfrac{1}{1 - \cfrac{z^2}{3 - \cfrac{z^2}{5 - \cfrac{z^2}{\ddots}}}}, \tag{20}$$

was in fact discovered earlier than Gauss by the Swiss-German mathematician Johann Lambert (1728–1777): Lambert developed (20) directly in order to prove that π is irrational, thereby solving a 2,000 year conjecture (it already appears in Aristotle [??]).

Lambert's argument is so elegant that we cannot resist presenting it here. Consider the function $f(z) = \frac{\tan z}{z}$. The key point is that, for each rational p/q, the value $f(p/q)$ is irrational. However, since $f(\pi/4) = 4/\pi$, then $4/\pi$ (hence π) *must* be irrational.

Now, the fact that $f(p/q)$ is irrational is indeed relatively easy to establish (nowadays), once (20) has been found. First, the convergents of Lambert's fraction, that start as

$$\frac{P_0}{Q_0} = \frac{1}{1}, \; \frac{P_1}{Q_1} = \frac{3}{3-z^2}, \; \frac{P_2}{Q_2} = \frac{15-z^2}{15-6z^2}, \; \frac{P_3}{Q_3} = \frac{105-10z^2}{105-45z^2+z^4}, \; \cdots$$

are given by the usual "three-term recurrence", $Q_n(z) = (2n+1)Q_{n-1}(z) - z^2 Q_{n-2}(z)$. Second, the determinant identity implies generally that

$$f(z) = \frac{P_N(z)}{Q_N(z)} + \sum_{n=N+1}^{\infty} \left(\frac{P_n(z)}{Q_n(z)} - \frac{P_{n-1}(z)}{Q_{n-1}(z)} \right) = \frac{P_N(z)}{Q_N(z)} + \sum_{n=N}^{\infty} \frac{z^{2n}}{Q_{n-1}(z)Q_n(z)}.$$

Finally, instantiate the last formula at some $z = p/q$. One estimates from the defining recurrence that the polynomials $Q_n(p/q)$ grow superexponentially fast, roughly like $2^n n!$; also, since the polynomial Q_n has integer coefficients and is of degree $\le (n+1)$, then $q^{n+1}Q_n(p/q)$ is an integer, and so is $q^{n+1}P_n(p/q)$. Thus, the expansion

$$f(\frac{p}{q}) = \frac{q^{N+1}P_N(p/q)}{q^{N+1}Q_N(p/q)} + \sum_{n=N+1}^{\infty} \frac{qp^{2n}}{q^{2n+1}Q_{n-1}(p/q)Q_n(p/q)}$$

is such that the first term, A_N/B_N, is a rational number that approximates[2] $f(p/q)$ with an error that is $o(1/B_N)$. Therefore, $f(p/q)$ is irrational.

It is somewhat unfortunate that Thomas Stieltjes [78] is mostly known to undegraduates for the Riemann-Stieltjes integral while this concept was only developed in the course of deeper researches concerning continued fractions. Amongst the many continued fraction gems left by Stieltjes, one finds several expansions of Laplace transforms of hyperbolic functions that evaluate to derivatives of the ψ function (itself the logarithmic derivative of the gamma function). An interesting specialization is

$$4\zeta(3) - 4 = \frac{1}{1} \; \underset{+}{} \; \frac{1^2 \cdot 2}{1} \; \underset{+}{} \; \frac{1 \cdot 2^2}{1} \; \underset{+}{} \; \frac{2^2 \cdot 3}{1} \; \underset{+}{} \; \frac{2 \cdot 3^2}{1} \; \underset{+}{} \; \cdots.$$

Some twenty years after Stieltjes but independently, Ramanujan [4, Ch. 12] obtained

$$\zeta(3) = 1 + \frac{1}{2 \cdot 2} \; \underset{+}{} \; \frac{1^3}{1} \; \underset{+}{} \; \frac{1^3}{6 \cdot 2} \; \underset{+}{} \; \frac{2^3}{1} \; \underset{+}{} \; \frac{2^3}{10 \cdot 2} \; \underset{+}{} \; \cdots.$$

The fractions of Stieltjes and Ramanujan exhibit slow convergence while the denominators of the approximants grow much too fast to allow for irrationality to be deduced. However, Apéry stunned a sceptical audience (see [61] for a vivid account) when, in June 1978, he announced a proof of the irrationality of $\zeta(3)$ based on

$$\zeta(3) = \frac{6}{5} \; \underset{-}{} \; \frac{1^6}{117} \; \underset{-}{} \; \frac{2^6}{535} \; \underset{-}{} \; \frac{3^6}{1436} \; \underset{-}{} \; \cdots,$$

[2] The A_N/B_N fail to be best approximants but not by too much. For instance, with $p/q = 7/10$, the quantity B_{100} is about 10^{291} while the error given by A_{100}/B_{100} is of the order of 10^{-411}.

where the denominators are the values of $34n^3 + 51n^2 + 27n + 5$.

Although an alternative proof of the irrationality of $\zeta(3)$ (due to Beukers and based on integrals of Legendre polynomials) is now known, it should be of interest to elucidate fully the circle of identities surrounding the fractions of Stieltjes, Ramanujan, and Apéry. A good reason is that it is yet to be determined that Catalan's constant,

$$G = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^2},$$

is irrational, while there exist for it several continued fraction representations due to Stieltjes and Ramanujan.

References

[1] L. Andersson and I. Vardi, *The dynamics of the Kasner billiard,* Preprint 2000.

[2] Archimedes, *Measurement of the Circle,* in, I. Thomas, *Greek Mathematical Works, Vol. I,* Loeb Classical Library **335**, Harvard University Press, Cambridge, MA, 1980.

[3] M. Beeler, R.W. Gosper, and R. Schroeppel, *HAKMEM,* A.I. Lab Memo # 239, M.I.T. 1972.

[4] B.C. Berndt, *Ramanujan's Notebooks, Part II,* Springer Verlag, New York, 1989.

[5] P. Billingsley, *Ergodic Theory and Information,* Wiley, New York 1965.

[6] C. Brezinski, *History of Continued Fractions and Padé Approximants,* vol. 12 of *Springer Series in Computational Mathematics,* Springer-Verlag, New York, 1991.

[7] R. Bruggeman, *Modular forms of varying weight, III,* J. Reine Angew. Math. **371** (1986), 181–198.

[8] J. Caveing, *L'irrationalité, Dans les mathématiques grecques jusqu'à Euclide,* Presses Universitaires du Septentrion, Paris, 1998.

[9] H. Cohen, *A Course in Computational Algebraic Number Theory,* Springer Verlag, New York, 1993.

[10] G. Cornell, J.H. Silverman, and G. Stevens (editors), *Modular Forms and Fermat's Last Theorem,* Springer–Verlag, New York, 1997.

[11] H.S.M. Coxeter and S.L. Greitzer, *Geometry Revisited,* Mathematical Association of America, 1967.

[12] H. Darmon, *A proof of the full Shimura–Taniyama–Weil conjecture is announced,* Notices of the A.M.S. **46** (1999), 1397–1401.

[13] H. Daudé, P. Flajolet and B. Vallée, *An average–case analysis of the Gaussian algorithm for lattice reduction,* Combinatorics, Probability, and Computing **6** (1997), 397–433.

[14] H. Davenport, *On certain exponential sums,* J. Reine Angew. Math. **169** (1932), 158–176.

[15] H. Davenport, *Multiplicative Number Theory,* Second Edition, Springer Verlag, New York, 1980.

[16] R. Dedekind, *Erläuterungen zu den vorstehenden Fragementen,* in Dedekind's "Gesammelte matematische Werke, Bd. I," Chelsea, New York, 1968), p. 159–172.

[17] H. Deligne, *La conjecture de Weil I,* Publ. I.H.E.S. **43** (1974), 273–307.

[18] N. Dershowitz and E.M. Reingold, *Calendrical Calculations,* Cambridge University Press, New York, 1997.

[19] J.M. Deshouillers and H. Iwaniec, *Kloosterman sums and Fourier coefficients of cusp forms,* Inv. Math. **70** (1982), 219–288.

[20] H.G. Diamond and J. Vaaler, *Estimates for Partial Sums of Continued Fraction Partial Quotients*, Pacific Journal of Mathematics **122** (1986), 73–82.

[21] J. Dixon, *The Number of Steps in the Euclidean Algorithm*, J. of Number Theory **2** (1970), 414–422.

[22] W. Duke, *Some old problems and new results about quadratic forms*, Notices of the A.M.S. **44** (1997), 190–196.

[23] Dutt (Romesh Cunder), *A History of Ancient India Based on the Sanskrit Literature*, Vishal Publishers, Delhi, 1972 (reprint of the original 1888 edition).

[24] B. Edixhoven, *Rational elliptic curves are modular [after Breuil, Conrad, Diamond and Taylor]*, Séminaire BOURBAKI, 1999–2000, n°871.

[25] ΕΥΚΛΕΙΔΟΥ ΣΤΟΙΧΕΙΑ, Euclide, Les Éléments, Texte grec et traduction française libre par G.J. Kayas, Editions du CNRS, Paris 1978.

[26] *Euclid Elementa*, edited by J.L. Heiberg and E.S. Stamatis, 5 vols, Teubner, Leipzig, 1969–77.

[27] Euclid, *The Thirteen Books of Euclid's Elements*, translated with introduction and commentary by T.L. Heath, Dover 1956.

[28] W. Feller, *An introduction to Probability Theory and its Applications, Volume II*, John Wiley& Sons, New York 1966.

[29] D.H. Fowler, *The Mathematics of Plato's Academy: a New Reconstruction*, Clarendon Press, Oxford 1990.

[30] J. Friedlander and H. Iwaniec, *The polynomial X^2+Y^4 captures its primes*, Annals of Math. **148** (1998), 945–1040.

[31] B.V. Gnedenko and A.N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison Wesley, Reading 1968.

[32] D. Goldfeld and P. Sarnak, *Sums of Kloosterman sums*, Invent. Math. **71** (1983), 243–250.

[33] R.W. Gosper, Electronic mail, September 23, 1992.

[34] Y. Guivarc'h and Y. Le Jan, *Asymptotic winding of the geodesic flow on modular surfaces and continuous fractions* , Ann. Scient. Éco. Norm. Sup. **26** (1993), 23–50.

[35] R.C. Gunning (notes by A. Brumer), *Lectures on Modular Forms*, Princeton University Press, Princeton, 1962.

[36] M. Hall, Jr., *On the sum and product of continued fractions*, Annals of Math. **48** (1947), 966–993.

[37] G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers*, Clarendon Press, Oxford, 1979.

[38] H. Heilbronn, *On the average length of a class of finite continued fractions*, in "Abhandlungen aus Zahlentheorie und Analysis," VEB Deutscher Verlag, Berlin 1968.

[39] L. Heinrich, *Rates of convergence in stable limit theorems for sums of exponentially ψ–mixing random variables with applications to metric theory of continued fractions*, Math. Nachr. **131** (1987), 149–165.

[40] P. Henrici, *Applied and Computational Complex Analysis, Vol. 2*, John Wiley, New York, 1974.

[41] D. Hensley, *The number of steps in the Euclidean algorithm*, J. Number Theory **49** (1994), 142–182.

[42] I.A. Ibragimov and Yu. V. Linnik, *Independent and stationary sequences of random variables*, Noordhoff, Groningen 1975.

[43] M. Kac, *Statistical Independence in Probability, Analysis and Number Theory,* Carus Math. Monographs **12**, Math. Association of America 1959.

[44] A. Ya. Khinchin, *Metrische Kettenbrucheprobleme*, Compositio Math. **1** (1935), 361–382.

[45] A. Ya. Khinchin, *Continued Fractions,* University of Chicago Press, Chicago 1964.

[46] A. Ya. Khinchin and P. Lévy, *Sur les lois stables,* C. R. Acad. Sci. Paris **202** (1936), 484-488.

[47] R. Kirby and P. Melvin, *Dedekind sums, μ–invariants and the signature cocycle,* Math. Ann. **299** (1994), 231–267.

[48] H.D. Kloosterman, *Asymptotische Formeln für die Fourierkoeffizienten ganzer Modulformen,* Abh. Math. Sem. Univ. Hamburg **5** (1927), 338–352.

[49] D.E. Knuth, *Notes on generalized Dedekind sums,* Acta Arith. **33** (1977), 297–325.

[50] D.E. Knuth, *The Art of Computer Programming, Vol. 2, Seminumerical Algorithms, 3rd Edition,* Addison–Wesley 1999.

[51] D.E. Knuth and A. Yao, *Analysis of the subtractive algorithm for greatest common divisors,* Proc. Nat. Acad. Sci. **72** (1975), 4720–4722.

[52] N.V. Kuznetsov, *The Petersson conjecture for cusp forms of weight zero and the Linnik conjecture. Sums of Kloosterman sums. (Russian),* Math. Sbornik (N.S.) **111(153)** (1980), 334–383.

[53] S. Levy, *Hackers: heroes of the computer revolution,* Doubleday, 1984.

[54] P. Liardet and P. Stambul, *Algebraic computations with continued fractions,* J. Number Theory **73** (1998), 92–121.

[55] Y.V. Linnik, *Additive problems and eigenvalues of modular operators,* I.C.M. Stockholm (1962), 270–284.

[56] W. Luo, Z. Rudnick, and P. Sarnak, *On Selberg's eigenvalue conjecture,* Geom. Funct. Anal. **5** (1995), 387–401.

[57] A.J. MacLeod, *High–accuracy numerical values in the Gauss–Kuzmin continued fraction problem,* Computers and Mathematics with Applications **26** (1993), 37–44.

[58] L.J. Mordell, *Lattice pointsin a tetrahedron and generalized Dedekind sums,* J. Indian Math. Soc. **15** (1951), 41–46.

[59] O. Perron, *Die Lehre von der Kettenbrüchen, Vol. 2,* Teubner, Leipzig, 1954.

[60] W. Philipp, *Limit Theorems for Sums of Partial Quotients of Continued Fractions,* Monat. Math. **105** (1988), 195–206.

[61] A. van der Poorten, *A proof that Euler missed … Apéry's proof of the irrationality of* $\zeta(3)$, Mathematical Intelligencer **1** (1979), 195–203.

[62] A. van der Poorten, *An introduction to continued fractions,* in "Diophantine Analysis (Kensington 1985)", London Math. Soc. Lecture Notes **109**, Cambridge University Press, Cambridge, 1986, p. 99–138.

[63] H. Rademacher, *Topics in analytic number theory,* Springer–Verlag, New York, 1973.

[64] H. Rademacher and E. Grosswald, *Dedekind Sums,* Carus Math. Monographs **16**, M.A.A., Washington, D.C., 1972.

[65] G.N. Raney, *On continued fractions and finite automata,* Math. Ann. **206** (1973), 265–283.

[66] B. Riemann, *Fragmente über die Grenzfälle der elliptischen Modulfunktionen,* in "Gesammelte mathematische Werke," based on the edition by H. Weber and R. Dedekind, Springer–Verlag, Berlin, 1990, p. 466–478.

[67] V.A. Rohlin, *Exact endomorphisms of Lebesgue spaces,* Izv. Akad. Nauk. SSSR **25**, 499–530.

[68] H. Salié, *Über die Kloostermannschen Summen $S(u,v;q)$,* Math. Z. **34** (1931), 91–109.

[69] P. Sarnak, *Some applications of modular forms,* Cambridge University Press, Cambridge 1990.

[70] R. Schoof, *Elliptic curves over finite fields and the computation of square roots mod p,* Math. Comp. **43** (1985), 483–494.

[71] A. Selberg, *On the estimation of Fourier coefficients of modular forms,* in "Theory of Numbers, Symposium on Recent Developments, Caltech 1963, Proc. Sympos. Pure Math. **8**, AMS, Providence, 1965, p. 1–15. Also in A. Selberg, *Collected Papers, Vol. 1,* Springer–Verlag, Berlin, 1989, p. 506–520.

[72] A. Selberg, *Reflections around the Ramanujan centenary,* in "Atle Selberg, Collected Papers, Vol. I," Springer Verlag, Berlin, 1989, 695–706.

[73] C. Series, *The modular function and continued fractions,* J. London Math. Soc. (2) **31** (1985), 69–80.

[74] J.–P. Serre, *Cours d'arithmétique, 2eme édition,* Presse Universitaires de France, Paris, 1977. English translation, *A Course in Arithmetic,* Springer–Verlag, New York, 1973.

[75] J.A. Serret and C. Hermite, J. de Math. Pures et Appl. **5** (1848), 12–15.

[76] J.O. Shallit, *Origins of the analysis of the Euclidean algorithm,* Historia Mathematica **21** (1994), 401–419.

[77] H.J. Smith, J. Reine Angew. Math. **50** (1855), 91–92.

[78] T.J. Stieltjes, *Œuvres Complètes,* edited by Gerrit van Dijk, Springer Verlag, Berlin, 1993.

[79] G. Tenenbaum, *Introduction à la Théorie Analytique et Probabiliste des Nombres, 2eme Édition,* Société Mathématique de France, Paris, 1995. English translation, *Introduction to Analytic and Probabilistic Number Theory,* Cambridge University Press, Cambridge, 1995.

[80] I. Vardi, *A relation between Dedekind sums and Kloosterman sums,* Duke Math. J. **55** (1987), 189–197.

[81] I. Vardi, *Computational Recreation in Mathematica,* Addison Wesley 1991.

[82] I. Vardi, *Dedekind sums have a limiting distribution,* Inter. Math. Research Notices (1993), 1–12, in Duke Math. J. **69** (1993). Duke Math. Journal.

[83] I. Vardi, *Code and Pseudo Code,* Mathematica Journal **6** Issue 2 (1996), 66–71.

[84] I. Vardi, *The St. Petersburg game and continued fractions,* C. R. de l'Académie des Sciences **324**, Série I (1997) 913–918.

[85] I. Vardi, *Archimedes' Cattle Problem,* American Math. Monthly. **105** (1998), 305–319.

[86] I. Vardi, *What is ancient mathematics?* Mathematical Intelligencer **21** (1999) no. 3, 38–47.

[87] J. Wainright and G.F.R. Ellis, *Dynamical Systems in Cosmology,* Cambridge University Press, Cambridge 1997.

[88] H.S. Wall, *Analytic Theory of Continued Fractions,* Chelsea, New York, 1948.

[89] A. Weil, *On some exponential sums,* Proc. Nat. Acad. Sci. **34** (1948), 204–207.

[90] A. Wiles, *Modular elliptic curves and Fermat's Last Theorem,* Annals of Math. **141** (1995), 443–551.

[91] D. Zagier, *Nombres de classes et fractions continues,* (Journées Arithmétiques de Bordeaux, 1974), Asterisque **24**–**25**, Soc. Math. France, Paris, 1975, p. 81–97.

[92] V.M. Zolotarev, *One–dimensional Stable Distributions,* Translations of Mathematical Monographs **65**, American Math. Soc., Providence 1986.