

An Informal Introduction to Continued Fractions

Euclid's method for the GCD,
square roots and Pell's equation,
the CFRAC factorisation algorithm,
sequences of convergents and errors,
rational approximation of irrationals,
Lagrange-Markoff spectrum,
Liouville's transcendental number theorem,
statistics of digits in 'general' numbers,
Thiele's curve fitting algorithm, and
functions $f(x)$ defined by continued fractions

John Coffey
Cheshire, UK

2011

This is an informal introduction to several aspects of continued fractions, using elementary maths and written in a narrative style at the level of the undergraduate student or interested amateur.

Continued fractions are fractions in which the denominator is an integer plus another fraction. For instance

$$2 + \frac{1}{3 + \frac{1}{5}} \quad \text{and} \quad \frac{1}{3 + \frac{1}{5 + \frac{1}{17}}}.$$

Each of these examples is readily evaluated by the elementary rules for fractions. The first represents the common fraction 37/16 and the second is 86/285.

This evaluation process can be reversed. Any given real number θ can be expressed as a continued fraction by continually repeating the process of extracting the integer part, a , to leave a remainder $\rho < 1$, taking the reciprocal of this remainder to be a new number $\theta_1 > 1$, extracting its integer part, and so on. Thus

$$\theta \equiv \theta_0 = a_0 + \rho_0, \quad \frac{1}{\rho_0} = \theta_1 = a_1 + \rho_1, \quad \frac{1}{\rho_1} = \theta_2 = a_2 + \rho_2, \quad \text{etc.}$$

At each level k , θ_k is the ' k^{th} complete quotient' and a_k is the ' k^{th} partial quotient'. If any remainder ρ_F is zero, the process naturally terminates with the final partial quotient a_F (F for 'final'). Such

terminating, finite continued fractions evaluate to ordinary fractions (rational numbers, \mathbb{Q}). On the other hand, non-terminating ones represent irrationals, which form the vast majority of the real numbers, \mathbb{R} . The continued fraction representation is a natural, intrinsic way to express any real θ , through recursively identifying its nested integer and fractional parts. This contrasts with the binary and decimal representations in which θ is repeatedly compared against the externally imposed reference numbers 2 and 10.

The iterative process of forming the continued fraction representation of any rational number is entirely equivalent to Euclid's algorithm for finding the greatest common divisor of two integers, these being the numerator and denominator of the corresponding fraction.

Truncating a continued fraction at some level k and evaluating the leading part as an ordinary fraction gives the ' k^{th} convergent', $C_k = p_k/q_k$. Because of the intrinsic representation property, the convergents of a continued fraction give closer rational approximations to the limiting value of the continued fraction than any other fraction with comparable denominator. The fact that convergents with quite low denominators can give accurate rational approximations to irrational numbers (such as $\pi \approx 22/7$) has been both of practical value in calculations for decades, and a tool for analysing the nature of real numbers, \mathbb{R} . However, continued fractions are cumbersome to manipulate – there is no ready way to add or multiply them.

After a brief introduction to the notation and essential arithmetic properties of continued fractions, the article elaborates on five major topics:

1. how to determine the continued fraction representation of a given rational (\mathbb{Q}) or real (\mathbb{R}), and conversely how to evaluate a given continued fraction as an ordinary fraction.
2. some practical applications of continued fractions to other aspects of maths,
3. the error in approximating a given real number by one of its convergents C_k and its implications for categorising \mathbb{R} ,
4. patterns of the partial quotients a_k and their the probability distribution for the generality of real numbers,
5. the representation of analytic functions $f(x)$ by generalised continued fractions.

The article deals with these topics in this order, though there is inevitably some overlap because the issues are interconnected. Also, I find it useful to keep the distinction between finite continued fractions $\in \mathbb{Q}$ and infinite fractions $\in \mathbb{R}, \notin \mathbb{Q}$. For this reason properties of finite continued fractions are generally dealt with before their infinite counterparts.

Part I opens by laying out in §1 some essential properties on which almost all of the rest of the article depends. Key topics are the recursion relation between convergents, the difference between adjacent convergents, and bounds on the error of the k^{th} convergent in approximating θ . Part 1 then considers Euclid's famous algorithm for finding the greater common divisor, g.c.d, (or highest common factor, h.c.f) of two integers. It examines representations of quadratic surds (*i.e.* those involving only square roots), which have continued fractions with a recurring sequence of partial quotients. Lagrange's method can be used to determine the continued fraction for cubics and some higher irrationals. Finally Part I introduces some generalisations from simple continued fractions.

Part II builds on Part I to explain three applications: 1) the calculation of logarithms, 2) Pell's equation $x^2 - Ny^2 = \pm k$ for integers $x, y, N > 0$ and k , which can be solved using the continued

fraction representation of \sqrt{N} , and 3) the CFRAC integer factorisation algorithm. Copious examples are given of each.

The convergence of the series C_k for some θ , and of the errors $\epsilon_k = |C_k - \theta|$, turns out to be a deep subject. Part III starts by looking at the symmetry group structure relevant to convergents C_k , and provides proof of their best approximation/best fit property, which is merely stated in Part I. Part III then analyses the rate of convergence of the series of convergents C_k . Some quadratic surds converge asymptotically precisely as a geometric series, and almost all continued fractions have a trend towards exponential convergence. The error ϵ_k is shown in many cases to depend on $1/q_k^2$, where q_k is the denominator. The study of errors in the convergents C_k leads to a classification of all reals in terms of their ability to be approximated efficiently by rationals. There is a hierarchy from rationals, to quadratic surds, to higher algebraic numbers and eventually to transcendentals. We look in §12 at the Lagrange/Markoff spectrum of bounds on error, where I hope the reader will find my account more simple and transparent than some in the literature. This explores some of the links between continued fractions and binary quadratic forms. §13 gives a proof of Liouville's theorem for the existence of transcendental numbers, and give some examples. Part III closes with a brief account of Cantor's pioneering work on transfinite numbers, some of which he developed using continued fractions.

Whereas Part I examined the recurring partial quotients a_k of quadratics (in which the a_k are constrained to particular values), Part IV uses probability theory to examine the patterns in the a_k for the generality of typical, 'generic' numbers. First we evaluate the continued fractions of sets of rationals p/q , $1 \leq p \leq q-1$ and study the patterns in $a_1, a_2, a_3, \text{ etc}$, counting the frequency at which $a_1, a_2 \dots$ equals 1 or 2 or 3, or \dots . Part 4 then moves on to reals θ and their representation as infinite continued fractions, and looks into the probability distribution of the values of the partial quotients for 'almost all' real numbers. Here we find a parallel with Benford's Law: the strange result for the distribution of leading digits in many naturally occurring data sets, namely that the digit 1 occurs most often. A thorough study of this deep topic would take us into the measure theory of probability and ergodic theory, but I just skate the surface. Nevertheless, some celebrated results in probability theory by Gauss, Kuzmin, Khinchin and Lévy are outlined.

Part V extends the concept of continued fractions to defining continuous functions $f(x)$ of a variable x . There is a strong link with infinite series, so questions of convergence arise. Some examples are $\tan x$, $\exp x$, and Gauss's hypergeometric function. The theory is also related to the approximation of functions of a variable x by a rational function of x , as developed by Padé and Thiele. We describe Thiele's algorithm for fitting a rational function to a given data set.

Continued fractions have a long history but are no longer taught in many maths courses, and the subject at undergraduate level has largely gone out of fashion. This introductory, non-rigorous account comes about through my own reading and explorations. So although parts of this personal account result from my own rediscovery or rewording of old truths, the mathematics is very well described in the literature. A major early treatise was by the great Leonard Euler, published in 1737. Amongst the illustrious names of mathematicians who have contributed are Wallis, Euler, Lagrange, Lambert, Gauss, Stieltjes and Ramanujan. The more I have read around this subject, the more books and articles I find which show that continued fractions are linked to many active areas of mathematical research such as Diophantine approximation and the analysis of \mathbb{R} , functional analysis, integer factorisation, measure theory and ergodic theory. I hope readers will be stimulated to explore this rich topic for themselves.

Index

Part I

Introductory numerical examples

Notation and terminology

Important ordering, recursion and difference relations

Introduction to error of approximation

Expressing a rational number as a continued fraction

Cutting sequence representation of a continued fraction Recurring continued fractions and quadratic surds

\sqrt{N} expressed as a recurring continued fraction

Special recursion sequences in \sqrt{N}

Continued fractions of cubics and other irrational numbers

Lagrange's method

Part II

Calculation of the logarithm of a given number

Pell's equation

The CFRAC factorisation algorithm

Part III

Symmetry groups related to convergents

Proofs of best fit and related properties of convergents.

Rate of convergence of quadratic continued fractions

Error estimates for continued fractions and the accurate approximation of irrationals

Hermite's limit on errors, and binary quadratic forms

Hurwitz theorem

Numerical evidence for error bounds

The Lagrange/Markoff spectrum of bounds on error

Markoff's equation and Markoff forms

Liouville's rational approximation theorem and related theorems on transcendental numbers (incomplete)

Cantor's studies on the cardinality of \mathbb{Q} , \mathbb{R} and \mathbb{R}^n .

Part IV (Incomplete)

Patterns in the sequence of partial quotients a_k for rationals.

Benford's law and the statistics of partial quotients a_k for real numbers.

Gauss-Kuzmin-Lévy distribution, Khinchin's constant, Lévy's constant

Part V (Incomplete)

Functions $f(x)$ defined by generalised continued fractions

Gauss's hypergeometric function

Thiele's algorithm for curve fitting with rational functions.

Some useful reference books

Many books on this topic are out of print. However Google Books has made some available to read free on the Internet, though several pages are always omitted.

1. Perhaps the most gentle introduction is the Open University booklet ‘Continued Fractions’ by Alan Best, Unit 7 of course M381 Number Theory and Mathematical Logic, 1996.
2. The classic ‘An Introduction to the Theory of Numbers’ by Hardy and Wright (Oxford Science Publications) has two chapters which derive many basic properties and build on them to develop transcendental numbers.
3. Several properties are clearly explained in ‘Approximation by Algebraic Numbers’ by Yann Bugeaud in the Cambridge Tracts in Mathematics series (2004).
4. Chapter 4 of ‘The Higher Arithmetic’ by H. Davenport in the Hutchinson University Library series is very clearly written and readable. Second Ed. 1962.
5. ‘Continued Fractions : Analytic Theory and Applications’ by W. B. Jones and W. J. Thron. A wide ranging textbook. Vol 11 in Encyclopedia of Mathematics and its Applications. Pub Addison-Wesley 1980, later by Cambridge University Press 1984.
6. ‘Continued Fractions’ by Andrew Rockett and Peter Szuesz is on Google Books. Pub. World Scientific 1992.
7. ‘Continued Fractions’ by Doug Hensley is also on Google Books. Pub. World Scientific 2006.
8. ‘Handbook of Continued Fractions for Special Functions’ by A.Cuyt, V.B. Petersen *et al.*. Pub Springer 2008. Available on GoogleBooks.
9. A major textbook which deals largely with algebraic analysis and convergence is ‘Continued Fractions’ by H. S. Wall, published by Van Nostrand, 1948. Available on GoogleBooks.
10. Claude Brezinski has written a comprehensive history of continued fractions and Padé approximations. Pub. Springer-Verlag 1980.
11. Ramanujan’s discoveries are explained in Part II of Bruce Berndt’s painstaking study of ‘Ramanujan’s Notebooks’ Pub. Springer-Verlag 1989.
12. ‘Making Transcendental Transparent’ by Edward Burger and Robert Tubbs, published Springer, is an excellent readable account of Liouville and Roth’s theorems and transcendental numbers.
13. Hans Riesel has a readable book ‘Prime numbers and computer methods of factorization’ pub. Birkhäuser, 1994, which explains CFRAC and other methods. It is available on Google Books.
14. ‘Gamma: Exploring Euler’s Constant’ by Julian Havil is a very readable book. Chapter 14 gives a clear account of Benford’s Law and its link to the probability of integers occurring as the partial quotients of a continued fraction. Princeton Univ. Press, 2003.
15. ‘The Calculus of Finite Differences’ by L M Milne-Thompson, Pub. Macmillan, 1933. Republished by American Mathematical Society 2000 and available in fragments on Google Books. Has a good description of linear finite difference equations and Thiele’s algorithm using reciprocal differences.

16. 'Probability and Measure' by Patrick Billingsley, 3rd edition, Pub. Wiley 1995. This is a thorough account of the title topic. §24 covers the ergodic theorem and its application to continued fractions.
17. Cantor's correspondence with Dedekind is described and annotated in 'Georg Cantor', a biography by Joseph Dauben, Harvard University Press, 1979. You can read some of Cantor's later papers in an English translation by Philip Jourdain, 1915, in 'Contributions of the founding of the theory of Transfinite Numbers', Dover, 1955. Jourdain's long introduction gives background on Cantor's use of continued fractions.
18. 'An Introduction to Diophantine Approximation' by John W. S. Cassels of Cambridge University is a rich monograph covering continued fractions, the Markoff spectrum, Roth's theorem and more. Publ. CUP, 1957.
19. There is an extensive literature on the Lagrange-Markoff Spectrum. The monograph of that title by Thomas Cusick and Mary Flahive is a modern summary, pub. 1989 American Mathematical Society. A summary which is as good as any is the thesis by Barbara Harzevoort for the University of Utrecht entitled 'Markoff Theory - A Geometric Approach', available on the internet at <http://igitur-archive.library.uu.nl/student-theses/2011-0330-200607/HarzevoortBarbaraMA2010.pdf>.

Part I

Evaluation and Calculation of Continued Fractions

1 Introduction: some essential facts

1.1 Introductory numerical examples

A continued fraction is formed when a mixed fraction (integer plus fraction) is written as the denominator of another fraction. You can continue nesting yet another fraction in the last denominator so far, so the depth of the nested fractions can be finite or continue infinitely. Finite ones evaluate to common fractions, that is to rational numbers \mathbb{Q} , whilst infinite ones represent irrational real numbers $\in \mathbb{R}$.

There are essentially two ‘directions’ for calculating with continued fractions:

Continued \rightarrow **ordinary**: given a continued fraction, to evaluate it as an ordinary, (‘common’ or ‘vulgar’) fraction, and conversely

Ordinary \rightarrow **continued**: given any ordinary number, to express it as a continued fraction.

These two modes of calculation are most readily illustrated with rational numbers because their continued fractions are finite.

1. Using basic school arithmetic, any finite continued fraction can be evaluated to an ordinary fraction. For example

$$1 + \frac{1}{2 + \frac{1}{3 + \frac{1}{4}}} = 1 + \frac{1}{2 + \frac{1}{13}} = 1 + \frac{1}{2 + \frac{4}{13}} = 1 + \frac{1}{\frac{30}{13}} = 1 + \frac{13}{30} = \frac{43}{30}.$$

2. You convert an ordinary fraction to a continued fraction by sequentially extracting the integer part from an improper, ‘top heavy’ fraction. If θ is any number in the range $0 < \theta < 1$, then $1/\theta > 1$. Subtract its integer component to leave another fraction < 1 , whose reciprocal in turn exceeds 1. The method is easily demonstrated by this example:

$$\frac{305}{131} = \frac{262 + 43}{131} = 2 + \frac{43}{131} = 2 + \frac{1}{\frac{131}{43}}$$

then

$$\frac{131}{43} = \frac{129 + 2}{43} = 3 + \frac{2}{43} = 3 + \frac{1}{\frac{43}{2}} = 3 + \frac{1}{21 + \frac{1}{2}}.$$

The $\frac{1}{2}$ at the last stage has reciprocal 2 with no remainder, so the sequence ends. We have obtained

$$\frac{305}{131} = 2 + \frac{1}{3 + \frac{1}{21 + \frac{1}{2}}}.$$

This procedure is essentially Euclid’s famous algorithm for the greatest common divisor (gcd) of the numerator and denominator of the given fraction. It is discussed in more detail in §2.

The essential point is that any real number θ can be expressed as a continued fraction by recursively extracting the integer part, a , to leave a remainder $\rho < 1$, taking the reciprocal of this remainder to be a new number $\theta_1 > 1$, extracting its integer part, and so on. Thus

$$\theta \equiv \theta_0 = a_0 + \rho_0, \quad \frac{1}{\rho_0} = \theta_1 = a_1 + \rho_1, \quad \frac{1}{\rho_1} = \theta_2 = a_2 + \rho_2, \text{ etc.} \quad (1.1)$$

The algorithm is easy to apply numerically, even with a hand calculator. At each level k , θ_k is called the ' k^{th} complete quotient' and a_k is the ' k^{th} partial quotient'. For any given real θ :

$$\begin{aligned} \theta &= a_0 + \frac{1}{\theta_1} = \frac{\theta_1 a_0 + 1}{\theta_1} \\ &= a_0 + \frac{1}{a_1 + \frac{1}{\theta_2}} = \frac{\theta_1(a_1 a_0 + 1) + a_0}{\theta_1 a_1 + 1}, \text{ etc.} \end{aligned}$$

A finite continued fraction results when one of the remainders ρ_F (F for 'final') is zero, in which case the last partial quotient is a_F .

The representation of any number as a continued fraction is unique¹. Unlike binary, decimals or hexadecimals, in which a fraction is represented as a series in powers of base 2, 10 or 16, continued fractions make no reference to any external base unit of counting. They are therefore an **intrinsic** representation.

1.2 Notation and terminology

The notation $\{a : b, c, d, \dots\}$ means

$$a + \frac{1}{b + \frac{1}{c + \frac{1}{d + \frac{1}{\dots}}}}$$

Some authors use different brackets and/or separators. An alternative notation for the above is used by other authors:

$$a + \frac{1}{b+} \frac{1}{c+} \frac{1}{d+} \dots$$

Yet another notation is

$$a + \frac{1|}{|b+} \frac{1|}{|c+} \frac{1|}{|d+} \dots$$

As mentioned above, the integers a , b , c , etc. are called *partial quotients*. Depending on context, I will denote the partial quotients by either a , b , c , etc. or by a_0 , a_1 , $a_2 \dots a_k$, etc. By convention none of these is negative. Moreover, none of b , c , d , etc. can be zero, though a (or a_0), the integer part of the given number θ , is zero if $\theta < 1$. Note that the numerator in each of the nested fractions is 1. Strictly, these are known as *simple* or *regular* continued fractions. There are various generalisations of continued fractions in which these numerators are other than +1, and these are mentioned in §3.5, 3.6 and 13.

¹There is a single exception for finite continued fractions such as $\{1: 2, 3, 4\}$, which can also be written as $\{1: 2, 3, 3, 1\}$ since $3 + \frac{1}{1} = 4$. Both evaluate to $\frac{43}{30}$. The implications of this are assessed in §2.3, Part I and §7.4, Part III.

Some special types extend to infinity by endless repetition of a finite sequence of partial quotients. Where a sequence such as b, c, d recurs indefinitely, the notation $\{a : \underline{b, c, d}\}$ means

$$a + \frac{1}{b + \frac{1}{c + \frac{1}{d + \frac{1}{b + \frac{1}{c + \frac{1}{d + \dots}}}}}}$$

When the continued fraction for a given θ is truncated at some partial quotient a_k (equivalent to setting the remainder ρ_k to zero), the corresponding common fraction is called the ' k^{th} convergent' of θ . Since truncation can be made after each partial quotient $a_0, a_1, a_2 \dots$ in turn, there is a sequence of convergents $C_0 = a_0, C_1 = a_0 + 1/a_1$, etc. Generically these are denoted $C_k = p_k/q_k$, where p_k is the numerator and q_k the denominator *as obtained directly from evaluation of the continued fraction without cancellation*. For instance, the successive convergents of $\{1: 2, 3, 4, 5, 6\}$, formed in effect by setting the successive remainders $\rho_k, k = 0, 1, 2, \dots$ to zero, are

$$C_0 = 1, \quad C_1 = \{1: 2\} = \frac{3}{2}, \quad C_2 = \{1: 2, 3\} = \frac{10}{7}, \quad C_3 = \frac{43}{30}, \quad C_4 = \frac{225}{157}, \quad C_5 = C_F = \theta = \frac{1393}{972}.$$

For a finite continued fraction, the final convergent C_F is clearly the value of θ itself. For an infinite fraction, θ must be understood as the limiting value of the infinite sequence of convergents.

We evaluate $\{1: 2, 3, 4, 5, 6\}$ by starting with the rightmost partial quotient and working leftwards. If evaluation is paused part way, the continued fraction has been split into two: the evaluated 'tail' and the more significant body still to be evaluated. For instance, $\{0: 5, 6\} = \frac{6}{31} = 0.193548$ so $\{1: 2, 3, 4, 5, 6\} = \{1: 2, 3, 4, \frac{6}{31}\}$. In general, $\theta = \{a_0 : a_1, a_2, \dots, a_k, a_{k+1}, \dots\} = \{a_0 : a_1, \dots, a_k + \rho_k\}$ will evaluate to a rational function of the tail ρ_k :

$$\frac{A\rho_k + C}{B\rho_k + D}$$

where $\frac{A}{B}$ and $\frac{C}{D}$ are convergents of θ . In fact $\frac{A}{B} = C_{k-1}$ and $\frac{C}{D} = C_k$. We will meet this important property again in §1.4.

We should expect that for $\theta = \{a_0 : a_1, a_2, a_3, \dots, a_k, \dots\}$.

1. the first (left-most, low k) partial quotients make the most significant contribution, whilst the high k ones 'fine tune' the value.
2. for any given index k , the larger the value of a_k , the larger the associated denominator and hence the smaller the contribution to θ .

On this basis we might expect $\{1: 1, 1, 1, 1, \dots\}$ to converge most slowly, and indeed it does.

1.3 An important ordering relation

As decimals the above sequence of convergents of $\{1: 2, 3, 4, 5, 6\}$ is

$$1.000 \quad 1.500 \quad 1.4286 \quad 1.43323 \quad 1.433121 \quad 1.433127.$$

Successive convergents are alternately lower then higher than the exact continued fraction – see Figure 1. For k even, $C_k < \theta$, and for k odd $C_k > \theta$. They therefore form two interwoven sequences,

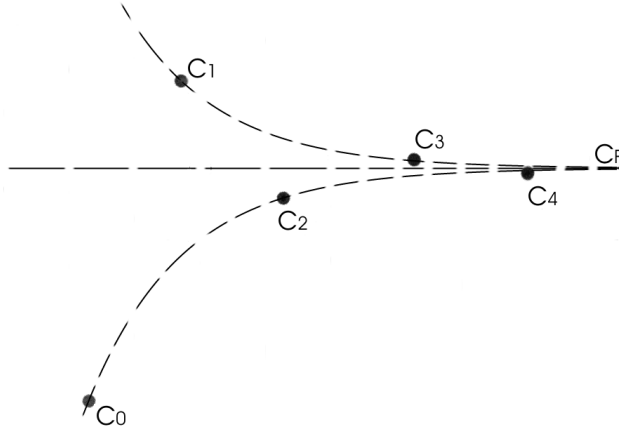


Figure 1: The convergents of θ form an alternating sequence

one converging to θ from below (k even) and the other from above. With any three successive convergents, the third has a value in between the first and the second. To see why, consider the fraction $\{a_0 : a_1, a_2, a_3, \dots, b, c, d, \dots\}$. The sub-fractions F formed by the partial quotients b, c and d are

$$F_b = b, \quad F_c = b + \frac{1}{c}, \quad F_d = b + \frac{1}{c + \frac{1}{d}}.$$

$$\text{Now } F_d = b + \frac{d}{cd + 1} \text{ which is less than } b + \frac{d}{cd} = F_c,$$

$$\text{so } F_b < F_d < F_c.$$

The effect is to give the convergent $\{a_0 : a_1, \dots, b, c, d\}$ a value between $\{a_0 : a_1, \dots, b\}$ and $\{a_0 : a_1, \dots, b, c\}$. A more thorough proof of these properties is given in §1.6.

We can express the same concept in a somewhat different way. In taking $\{1: 2, 3\}$ as an approximation to $\theta = \{1: 2, 3, 4, 5, 6\}$, we are saying that θ lies in the semi-open interval I_2 between $\{1: 2, 3.000\dots\}$ and $\{1: 2, 3.999\dots\} = \{1: 2, 4\}$. The next convergent, $\{1: 2, 3, 4\}$, locates θ within the interval I_3 from $\{1: 2, 3, 4\}$ to $\{1: 2, 3, 5\}$. I_3 is enclosed within I_2 . In this way, the convergents of θ define a set of nested intervals, each wholly contained within the previous. In this way the continued fraction algorithm defines a unique partitioning of the real number line.

It is possible to say of two continued fractions which represents the larger number by comparing their sequences of partial quotients. Find the lowest k at which the respective sequences differ. The sign $<$ or $>$ depends on whether k is even or odd. As an example for $k = 2$ (even), $\{4: 3, 2\} = 4.28571 < \{4: 3, 3\} = 4.30$, and note that a_2 is 2 in the lesser fraction and 3 in the greater. The opposite behaviour occurs for $k = 3$ (odd), as with $\{4: 3, 2, 5\} = 4.28947$, compared with $\{4: 3, 2, 6\} = 4.28888$. In general

$$\{a_0 : a_1, a_2, \dots, a_{k-1}, B\} < \{a_0 : a_1, a_2, \dots, a_{k-1}, C\} \text{ if } k \text{ is even and } B < C,$$

$$\{a_0 : a_1, a_2, \dots, a_{k-1}, B\} < \{a_0 : a_1, a_2, \dots, a_{k-1}, C\} \text{ if } k \text{ is odd and } C < B.$$

A consequence of this is that if only two values are used as the the partial quotients of some θ , $b > s$ ('big' > 'small'), the largest value of θ that can be represented is $\{b : \underline{s}, b\}$, and the smallest is $\{s : \underline{b}, s\}$, or possibly $\{0 : \underline{b}, s\}$.

1.4 An important recursion relation

An unhelpful feature of continued fractions is that there is little arithmetic that you can do with them directly. You cannot add or multiply them without first converting them to ordinary fractions. There is no obvious relation between the continued fraction for n and any of its multiples. For instance,

$$\frac{7}{23} = \{0 : 3, 3, 1, 1\}, \quad \frac{14}{23} = \{0 : 1, 1, 1, 1, 3, 1\}, \quad \frac{21}{23} = \{0 : 1, 10, 1, 1\}, \quad \frac{28}{23} = \{1 : 4, 1, 1, 2\}.$$

Moreover, to evaluate the continued fraction in the example of §1.1 we have had to start at the right hand end of the $\{1: 2, 3, 4, 5, 6\}$ sequence, and deal with the least significant partial quotient first. We now show that there is a way to start at the left end, with the most significant partial quotients.

The recursive algorithm Eq 1.1 for converting an ordinary fraction to a continued one imparts recursion relations which link the numerators p_k and the denominators q_k of adjacent convergents. Consider the first few convergents of $\{a_0 : a_1, a_2, a_3, \dots\}$:

$$\begin{aligned} C_0 &= \frac{p_0}{q_0} = \frac{a_0}{1}, & C_1 &= \frac{p_1}{q_1} = a_0 + \frac{1}{a_1} = \frac{a_1 a_0 + 1}{a_1}, \\ C_2 &= \frac{p_2}{q_2} = a_0 + \frac{1}{a_1 + \frac{1}{a_2}} = \frac{a_0 a_1 a_2 + a_0 + a_2}{a_2 a_1 + 1} = \frac{a_2(a_1 a_0 + 1) + a_0}{a_2 a_1 + 1} = \frac{a_2 p_1 + p_0}{a_2 q_1 + q_0}. \\ C_3 &= \frac{p_3}{q_3} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3}}} = a_0 + \frac{a_3 a_2 + 1}{a_1 a_2 a_3 + a_1 + a_3} = \frac{[(a_0 a_1 + 1) a_2 + a_0] a_3 + a_0 a_1 + 1}{(a_1 a_2 + 1) a_3 + a_1} = \frac{a_3 p_2 + p_1}{a_3 q_2 + q_1}. \end{aligned}$$

You can see the pattern. Three successive numerators and denominators are linked by

$$p_{k+1} = a_{k+1} p_k + p_{k-1} \quad \text{and} \quad q_{k+1} = a_{k+1} q_k + q_{k-1}. \quad (1.2)$$

These are probably the most important relations in this article. First, they show how p_k and q_k both increase rapidly with k ; in all cases $p_k < p_{k+1}$, $q_k < q_{k+1}$. Second, they allow evaluation of continued fractions from the left hand (most significant) partial quotient, finding higher convergents from the previous two. This makes it easy to tabulate convergents on a computer spreadsheet. (Of course, you must use the convergents exactly as calculated and not equivalent fractions from which common factors have been cancelled.) Each of these relations is a linear difference equation. For certain types of fraction, in which the partial quotients a_k are constant, it can be solved by well established techniques and yield explicit, closed form expressions for the p_k or q_k . We will investigate this in Part III in the context of quantifying the rate of convergence of the series of convergents.

Here is a proof of Eq 1.2 by induction on the index k . Let $\mathcal{H}(k)$ denote the hypothesis that $p_k = a_k p_{k-1} + p_{k-2}$. As base cases, the direct calculations above have shown that $\mathcal{H}(2)$ and $\mathcal{H}(3)$ are true. Our aim now is to show that the assumed truth of $\mathcal{H}(k)$ necessarily implies the truth of $\mathcal{H}(k+1)$. In general

$$\frac{p_k}{q_k} = a_0 + \frac{1}{\frac{p_{k-1}^*}{q_{k-1}^*}} = \frac{a_0 p_{k-1}^* + q_{k-1}^*}{p_{k-1}^*}$$

where the superscript $*$ denotes the operation of incrementing every index by 1 in p_k and q_k . Since there has been no cancellation,

$$\text{a) } q_k = p_{k-1}^* \quad \text{and} \quad \text{b) } p_k = a_0 p_{k-1}^* + q_{k-1}^*.$$

If we had instead evaluated p_{k+1}/q_{k+1} , we would have

$$\text{c) } q_{k+1} = p_k^* \quad \text{and} \quad \text{d) } p_{k+1} = a_0 p_k^* + q_k^*.$$

Increment by 1 all indices in the formula $\mathcal{H}(k) : p_k = a_k p_{k-1} + p_{k-2}$ (this is just a matter of relabelling the a_k), and substitute relation a):

$$\text{e) } p_k^* = a_{k+1} p_{k-1}^* + p_{k-2}^* \quad \rightarrow \quad \text{f) } q_{k+1} = a_{k+1} q_k + q_{k-1}.$$

f) is the denominator relation in Eq 1.2. Its derivation shows that the assumption of a recursion relation on the p_k implies an equivalent relation on the q_k . Next, again relabel indices by applying the $*$ operation to f) to obtain $q_k^* = a_{k+1} q_{k-1}^* + q_{k-2}^*$. Substitute this and e) into d), and use b):

$$\begin{aligned} p_{k+1} &= a_0 (a_{k+1} p_{k-1}^* + p_{k-2}^*) + a_{k+1} q_{k-1}^* + q_{k-2}^* \\ &= a_{k+1} (a_0 p_{k-1}^* + q_{k-1}^*) + (a_0 p_{k-2}^* + q_{k-2}^*) \\ &= a_{k+1} p_k + p_{k-1}. \end{aligned}$$

This is the formula expressing $\mathcal{H}(k+1)$ and so concludes the proof of Eq 1.2. In their book, page 130, Hardy and Wright give a shorter and more elegant proof.

I emphasise that in using Eq 1.2 to calculate the series of convergents C_k , it is necessary to use the p_k and q_k exactly as calculated from Eq 1.2; cancelling any common factors between p_k and q_k will cause errors in all higher convergents.

Finally, recall from §1.2 that if the partial remainder ρ_{k+1} is not set to zero, we retain the exact representation of θ . Then the partial quotient a_{k+1} in Eq 1.2 is replaced by the complete quotient θ_{k+1} to give

$$\theta \equiv \theta_0 = \frac{\theta_{k+1} p_k + p_{k-1}}{\theta_{k+1} q_k + q_{k-1}}. \quad (1.3)$$

For example, whilst the fraction $\theta = 1393/972 = \{1 : 2, 3, 4, 5, 6\}$ has convergent $C_3 = \{1 : 2, 3, 4\} = 43/30 = \frac{4 \times 10 + 3}{4 \times 7 + 2}$ (using Eq 1.2 with the previous convergents), θ is expressed exactly by $\frac{4.1935 \dots \times 10 + 3}{4.1935 \dots \times 7 + 2}$, where 4.1935... is the complete quotient θ_3 . Also, from the definition of θ_{k+1} in Eq 1.1, bear in mind that θ_{k+1} itself has the continued fraction $\{a_{k+1} : a_{k+2}, a_{k+3}, \dots\}$.

1.4.1 A note on matrix notation

The mathematician Louis Milne-Thomson has promoted writing the recursion relations Eq 1.2 as a matrix multiplication. We shall refer to this in places later in this article. He observed that

$$(a_k \quad 1) \begin{pmatrix} p_{k-1} & q_{k-1} \\ p_{k-2} & q_{k-2} \end{pmatrix} = (a_k p_{k-1} + p_{k-2} \quad a_k q_{k-1} + q_{k-2}) = (p_k \quad q_k).$$

There is a disadvantage here because the row matrix $(a_k \quad 1)$, not being square, cannot be multiplied by row matrices for other values of k . The problem is overcome by adding another row which merely reproduces the values of p_{k-1} and q_{k-1} :

$$\begin{pmatrix} a_k & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} p_{k-1} & q_{k-1} \\ p_{k-2} & q_{k-2} \end{pmatrix} = \begin{pmatrix} a_k p_{k-1} + p_{k-2} & a_k q_{k-1} + q_{k-2} \\ p_{k-1} & q_{k-1} \end{pmatrix} = \begin{pmatrix} p_k & q_k \\ p_{k-1} & q_{k-1} \end{pmatrix}. \quad (1.4a)$$

Taking $p_0 = a_0$, $q_0 = 1$, the k^{th} convergent is obtained by repeated matrix multiplication (with operations on the right being performed first). For example

$$\begin{pmatrix} a_3 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} p_3 & q_3 \\ p_2 & q_2 \end{pmatrix}.$$

It is sometimes more useful to have the p_k, q_k of the convergents written as columns rather than rows. In this case multiplication by the a_k matrices (which are symmetrical) is on the right as follows:

$$\begin{pmatrix} p_{k-1} & p_{k-2} \\ q_{k-1} & q_{k-2} \end{pmatrix} \begin{pmatrix} a_k & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} a_k p_{k-1} + p_{k-2} & p_{k-1} \\ a_k q_{k-1} + q_{k-2} & q_{k-1} \end{pmatrix} = \begin{pmatrix} p_k & p_{k-1} \\ q_k & q_{k-1} \end{pmatrix}. \quad (1.4b)$$

1.5 An important difference relation

A relation which is probably equally important and far reaching as Eq 1.2 is the difference between adjacent convergents, $\Delta_k = C_k - C_{k-1}$. Using the working in §1.4 which led to Eq 1.2,

$$\begin{aligned} \Delta_1 &= \frac{p_1}{q_1} - \frac{p_0}{q_0} = \frac{a_1 a_0 + 1}{a_1} - \frac{a_0}{1} = \frac{1}{a_1} = \frac{1}{q_0 q_1} \\ \Delta_2 &= \frac{p_2}{q_2} - \frac{p_1}{q_1} = \frac{p_2 q_1 - p_1 q_2}{q_2 q_1} = \frac{p_0 q_1 - p_1 q_0}{q_2 q_1} = \frac{a_0 a_1 - (a_1 a_0 + 1) \cdot 1}{q_2 q_1} = \frac{-1}{q_2 q_1} \end{aligned}$$

where terms $a_2 p_1 q_1$ have cancelled from the numerator. In general

$$\Delta_k = \frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} = \frac{p_k q_{k-1} - p_{k-1} q_k}{q_k q_{k-1}}.$$

The numerator is evaluated by using recursion relations Eq 1.2:

$$\begin{aligned} p_k q_{k-1} - p_{k-1} q_k &= (a_k p_{k-1} + p_{k-2}) q_{k-1} - p_{k-1} (a_k q_{k-1} + q_{k-2}) = p_{k-2} q_{k-1} - p_{k-1} q_{k-2} \\ &= p_{k-2} (a_{k-1} q_{k-2} + q_{k-3}) - (a_{k-1} p_{k-2} + p_{k-3}) q_{k-2} = p_{k-2} q_{k-3} - p_{k-3} q_{k-2}. \end{aligned}$$

Thus, reducing all the indices in the numerator by 2 leaves the value unchanged. Therefore all numerators with even values of k must have the same value, and similarly for odd. Since Δ_1 has a + sign and Δ_2 a - sign, in general

$$p_k q_{k-1} - p_{k-1} q_k = (-1)^{k-1} \quad (1.5)$$

$$\Delta_k \equiv C_k - C_{k-1} = \frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} = \frac{(-1)^{k-1}}{q_k q_{k-1}} \quad (1.6)$$

The alternating signs are entirely consistent with neighbouring convergents lying on opposite sides of θ on the real number line. Cast in the matrix notation of §1.4.1, Eq 1.5 states that the determinant of the matrix in Eq 1.4 is alternately -1 then $+1$ as k steps up from 0. We will make extensive use of Eqs 1.5 and 1.6 in subsequent sections; in particular, in proving the best rational approximation property of convergents.

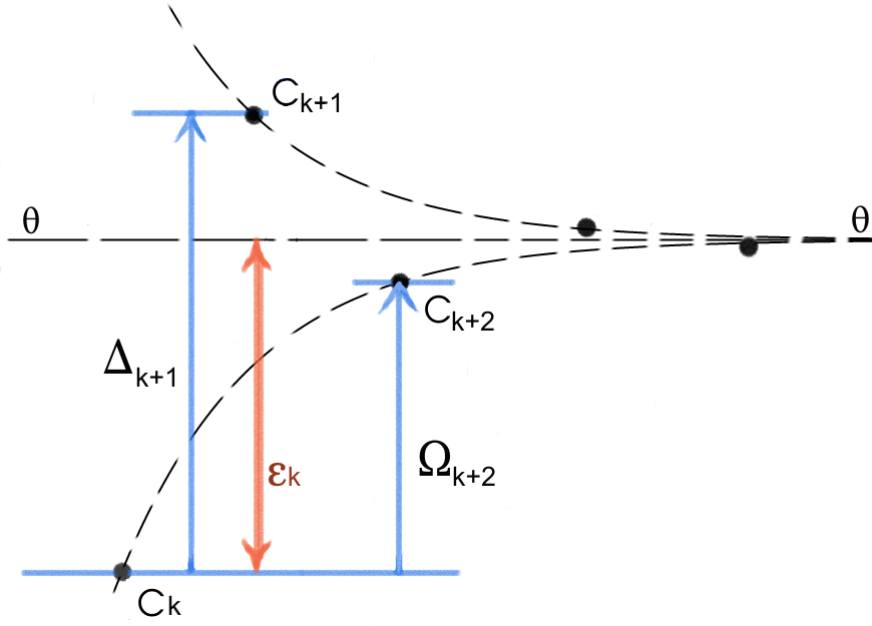


Figure 2: Illustrating Δ , ϵ and Ω as defined with respect to C_k

There is a formula similar to Eq (1.6) for the next-neighbour difference Ω

$$\Omega_k = C_k - C_{k-2} = \Delta_k + \Delta_{k-1} = \frac{(-1)^{k-1}(q_{k-2} - q_k)}{q_k q_{k-1} q_{k-2}} = \frac{a_k (-1)^k}{q_k q_{k-2}}, \quad (1.7)$$

where the recursion relation Eq 1.2 has been used to pull a factor q_{k-1} out of the numerator. The meanings of Δ and Ω are illustrated in Figure 2. Also shown is the error ϵ , described in the next subsection. Both Δ and Ω are signed quantities. Figure 2 corresponds to k even.

1.6 The ratio of adjacent convergents

Eqs 1.2 and 1.5 allow a more thorough proof of the convergence of the series of even and odd convergents than that given in §1.3 above. They can be used to find the ratio of two adjacent convergents, C_k and C_{k+1} , of $\theta = \{a_0 : a_1, a_2, \dots, a_k, \dots\}$.

$$\frac{C_k}{C_{k+1}} = \frac{p_k q_{k+1}}{q_k p_{k+1}} = \frac{p_k [a_{k+1} q_k + q_{k-1}]}{q_k [a_{k+1} p_k + p_{k-1}]} = \frac{a_{k+1} p_k q_k + p_k q_{k-1}}{a_{k+1} p_k q_k + p_{k-1} q_k}.$$

Use Eq 1.5 to replace $p_{k-1} q_k$ in the denominator:

$$\frac{C_k}{C_{k+1}} = \frac{a_{k+1} p_k q_k + p_k q_{k-1}}{a_{k+1} p_k q_k + p_k q_{k-1} + (-1)^k} = \frac{S_k}{S_k + (-1)^k}$$

where S_k is some polynomial in the first $k+1$ partial quotients. So the numerator and denominator of this ratio differ by precisely $+1$ or -1 , making the ratio alternately greater and less than 1. Moreover, S_k greatly exceeds the product of all squared partial quotients $\prod a_j^2$, $1 \leq j \leq k$, so increases rapidly and without limit with always $S_{k+1} > S_k$. Taking pairwise ratios of three consecutive convergents, C_{k-1}/C_k and C_k/C_{k+1} , proves that C_{k+1} lies between C_{k-1} and C_k . Also the ratio of adjacent convergents tends to 1 as $k \rightarrow \infty$, sandwiching the limiting value θ .

1.7 Error and ‘best fit’ property of convergents

The odd and even subsequences of convergents always converge monotonically to θ , so each C_k is an approximation to θ characterised by the size of its denominator, q_k . The error in C_k is an unsigned quantity defined by

$$\epsilon_k \equiv |C_k - \theta| .$$

and is illustrated in Figure 2. It decreases rapidly with increasing k . A numerical illustration for a typical continued fraction is given in Table 1.

Table 1: Example of convergents and errors

k	a_k	p_k	q_k	C_k	ϵ_k
0	6	6	1	6	0.214
1	4	25	4	6.25	0.036
2	1	31	5	6.20	0.014
3	2	87	14	6.21429	2.92E-04
4	17	1510	243	6.21399	2.06E-06
5	8	12167	1958	6.213993871	4.34E-08
6	5	62345	10033	6.213993820	7.50E-09
7	1	74512	11991	6.21399382870	8.10E-10
8	7	583929	93970	6.21399382782	7.75E-11
9	1	658441	105961	6.213993827917819	2.29E-11
10	3	2559252	411853	6.213993827894904	0

Much of this article discusses the error in the convergents – Part III is devoted to the topic. By using Eq 1.1, 1.2 and 1.3 we can derive an expression for the error ϵ_k as follows:

$$\begin{aligned} \epsilon_k &= \left| \frac{a_k p_{k-1} + p_{k-2}}{a_k q_{k-1} + q_{k-2}} - \frac{\theta_k p_{k-1} + p_{k-2}}{\theta_k q_{k-1} + q_{k-2}} \right| \\ &= \frac{\theta_k - a_k}{q_k (\theta_k q_{k-1} + q_k)} \\ &= \frac{\rho_k}{q_k (q_k + \rho_k q_{k-1})} \end{aligned} \tag{1.8}$$

This is an important exact formula, but of course the remainder ρ_k is generally not known. Much of the analysis will involve finding several practical estimates of ϵ_k , or inequalities which give upper or lower bounds on ϵ_k . In general, greater accuracy and tighter bounds on error imply more complicated formulae. A first approximation comes from noting that $\rho_k < 1$ and $q_{k-1} < q_k$. Therefore

$$\epsilon_k \approx \frac{\rho_k}{q_k^2} \quad \text{or} \quad \epsilon_k q_k^2 \approx \rho_k .$$

Much of the discussion of errors will normalise ϵ_k to $1/q_k^2$.

Consider now the various estimators of error listed in Table 2, which correspond to the continued fraction in Table 1. These illustrate the following statements which can be made at this stage regarding bounds on ϵ_k . The values in the columns in Table 2 follow the definitions below.

1. The simplest upper bound is $\epsilon_k < |\Delta_{k+1}|$ because

$$|\Delta_{k+1}| = |C_{k+1} - C_k| > |\theta - C_k| = \epsilon_k .$$

Table 2: Comparison of error estimates for example in Table 1

ϵ_k	$ \Delta_{k+1} $	$ \Delta_{k+1} /2$	$1/q_k^2$	$1/2q_k^2$	S_Δ	$S_{\Delta/2}$	S_{q^2}	$S_{q^2/2}$
0.214	0.250	0.125	1	0.5	1	-1	1	1
0.036	0.050	0.025	0.0625	0.0313	1	-1	1	-1
0.014	0.014	7.14E-03	0.0400	0.0200	1	-1	1	1
2.92E-04	2.94E-04	1.47E-04	5.10E-03	2.55E-03	1	-1	1	1
2.06E-06	2.10E-06	1.05E-06	1.69E-05	8.47E-06	1	-1	1	1
4.34E-08	5.09E-08	2.55E-08	2.61E-07	1.30E-07	1	-1	1	1
7.50E-09	8.31E-09	4.16E-09	9.93E-09	4.97E-09	1	-1	1	-1
8.10E-10	8.87E-10	4.44E-10	6.95E-09	3.48E-09	1	-1	1	1
7.75E-11	1.00E-10	5.02E-11	1.13E-10	5.66E-11	1	-1	1	-1
2.29E-11	2.29E-11	1.15E-11	8.91E-11	4.45E-11	0	-1	1	1
0			5.90E-12	2.95E-12			1	1

The column labelled S_Δ in Table 2 lists the sign function (+1, 0 or -1) for the difference $|\Delta_{k+1}| - \epsilon_k$, as a way of emphasising this inequality.

- To find a simple lower bound, θ lies between C_k and C_{k+1} with C_{k+1} always the better approximation. So θ lies between the mean $(C_{k+1} + C_k)/2$ and C_{k+1} . For k even

$$C_{k+1} > \theta > \frac{1}{2}(C_{k+1} + C_k) \quad \text{so} \quad C_{k+1} - C_k > \theta - C_k > \frac{1}{2}(C_{k+1} - C_k),$$

and similarly for k odd. Therefore $\frac{1}{2}|\Delta_{k+1}| < \epsilon_k$. In Table 2 the column of signs for $|\Delta_{k+1}|/2 - \epsilon_k$ is labelled $S_{\Delta/2}$.

- Another lower bound is that $|\Omega_{k+2}| < \epsilon_k$ (see Figure 2).
- These upper and lower bounds can be expressed in terms of the denominators rather than the differences Δ_k . This has the advantage of expressing the error in terms only of C_k and not both C_k and C_{k+1} . Use Eq 1.6 then the recursion relation Eq 1.2:

$$|\Delta_{k+1}| = \frac{1}{q_{k+1}q_k} = \frac{1}{(a_{k+1}q_k + q_{k-1})q_k} < \frac{1}{a_{k+1}q_k^2} \leq \frac{1}{q_k^2}. \quad (1.9)$$

since all $a_k \geq 1$. From 1) above,

$$\epsilon_k < \frac{1}{q_k^2}. \quad (1.10)$$

The column of signs for $1/q_k^2 - \epsilon_k$ is labelled S_{q^2} and is positive in all cases. This important statement is true for *all* convergents. It can be derived from Eq 1.8 by setting ρ_k in the numerator to 1 (supremum), and to 0 in the denominator (lowest possible value).

- Bound 2) above is $\frac{1}{2}|\Delta_{k+1}| < \epsilon_k$, whilst Eq 1.10 gives $\frac{1}{2}|\Delta_{k+1}| < 1/(2q_k^2)$. This leaves ambiguity as to how ϵ_k compares with $1/2q_k^2$. The example in Table 2 illustrates that for some convergents $\epsilon_k < 1/(2q_k^2)$ while for others it is greater, as emphasised by the variable sign in the last column.

Item 5) can be explained as follows. The following inequality holds for any two real numbers u and v :

$$2(u^2 + v^2) = (u+v)^2 + (u-v)^2 > (u+v)^2 - (u-v)^2 = 4uv.$$

Let $u = 1/(2q_k)$, $v = 1/(2q_{k+1})$ and use Eq (1.6)

$$\frac{1}{2q_k^2} + \frac{1}{2q_{k+1}^2} > \frac{1}{q_k q_{k+1}} = |\Delta_{k+1}|.$$

But θ lies between C_k and C_{k+1} , so $|\Delta_{k+1}| = |C_{k+1} - \theta| + |\theta - C_k|$. So

$$\epsilon_k + \epsilon_{k+1} < \frac{1}{2q_k^2} + \frac{1}{2q_{k+1}^2}. \quad (1.11a)$$

If $\epsilon_{k+1} > \frac{1}{2q_{k+1}^2}$, then $\epsilon_k < \frac{1}{2q_k^2}$ and *vice versa*.

(1.11b)

In words, at least one convergent C_k of any adjacent pair must have an absolute error less than $1/2q_k^2$. The error in the adjacent convergent $C_{k\pm 1}$ can be greater than $1/2q_{k\pm 1}^2$, but within the constraint that it is less than $1/q_{k\pm 1}^2$. Observe that in Table 2, last column, any values of -1 (corresponding to $\epsilon_k > 1/2q_k^2$) are next to values of $+1$; two adjacent -1 values never occur.

Further analysis, as given by Hardy and Wright and proved in Part III, concludes that the converse of Eq 1.11 is true: namely, that for any real number θ , if a rational number u/v satisfies

$$\left| \frac{u}{v} - \theta \right| < \frac{1}{2v^2}, \quad (1.12)$$

then u/v is a convergent in the continued fraction representation of θ . In other words if any fraction u/v comes within $1/(2v^2)$ of a real number θ , then u/v must be a convergent of θ . Moreover, there is an infinite number of fractions with this property, being the infinite set of convergents. This is one of the most significant results in this article.

Here is an imaginary situation to illustrate the significance of Eq 1.11. Suppose you are the contestant in a TV quiz show, with big money stakes, called ‘Guess My Number’. The quiz master poses this challenge :

‘I am going to show you a long decimal number, θ . I want you to tell me a fraction, p/q , which approximates my number so closely that the error is less than $1/2q^2$. You have only three guesses. The given number is (*pause for suspense*) ... ’.

He holds up a board on which is written $\theta = 4 \cdot 1032159733$.

What a challenge! Under the pressure of the audience and the studio lights you blurt out the first approximation which comes to your head, the truncated decimal $4 \cdot 103$. The quiz master pounces. ‘You have said $4103/1000$. The error is 2.16×10^{-4} . But your target error is $1/(2 \times 1000^2) = 5 \times 10^{-7}$. No, you have failed at your first attempt. You have two guesses left.’

What to do? Pictured in terms of the real number line, the denominator q defines a fineness to the scale of fractions $1/v, 2/v, \dots, (u-1)/v, u/v, (u+1)/v, \dots$. So if you make the denominator v larger, you will get a closer spacing of adjacent fractions and precision will improve proportional to v , allowing you to choose a closer approximation. But on the down side, the error target will shrink even faster, as $1/v^2$. On the other hand if you make the denominator v smaller, the target will become easier to hit, but the gradations between multiples of $1/v$ will then be so coarse that not one is sufficiently close to θ , even with the target error relaxed.

What you really need is a way of identifying a denominator, perhaps not very different from 1000, which places the grid of fractions in such a way that one u/v value lies especially close to θ . These (as you know) are the convergents of θ . The quiz master's target is so narrow that only a convergent can achieve it, and indeed, from Eq 1.10, perhaps only one convergent in any adjacent pair will. But you have two guesses left. The continued fraction for θ is $\{4: 9,1,2,4,1,3,2,2,1,1,\dots\}$ and the convergents with denominator either side of 1000 are $2425/591$ and $5486/1337$. You will win with at least one of these. At home afterwards, you breathe a sigh of relief and check using your computer whether there are any other fractions with $500 < v < 1500$ for which $|u/v - \theta| < 1/2v^2$. There aren't any.

This is what is meant by saying that continued fractions yield the most sharp, most efficient approximations to any given real number θ . The k^{th} convergent $C_k = p_k/q_k$ is the closest possible approximation to θ of all fractions with denominator $\leq q_k$. To obtain a better approximation, you need a larger denominator. However, the larger the denominator, the less efficient the approximation is considered to be – perhaps because, in days before electronic calculators, they were more tedious and error-prone for hand calculations. Older readers may remember doing hand calculations at school using $\pi \approx 22/7 = 3\frac{1}{7}$. This is in fact the first convergent of π . It is a better approximation than $3 \cdot 1 = 3\frac{1}{10}$, the equivalent approximation using decimals, even though 10 is larger than 7. §8 in Part III looks more deeply into errors ϵ_k , giving proofs of important relations.

Table 3: Close errors bounds, Eq 1.12, for example in Tables 1 & 2

ϵ_k	lower	upper
0.214	0.167	0.250
0.036	0.021	0.063
0.014	0.010	0.020
2.92E-04	2.69E-04	3.00E-04
2.06E-06	1.69E-06	2.12E-06
4.34E-08	3.73E-08	5.22E-08
7.50E-09	3.31E-09	9.93E-09
8.10E-10	7.73E-10	9.94E-10
7.75E-11	3.77E-11	1.13E-10
2.29E-11	1.78E-11	2.97E-11
0	2.95E-12	

To conclude this Section, let us determine close upper and lower bounds on error, based on the analysis so far. These bounds follow from Eq 1.7 and Eq 1.9:

$$\epsilon_k > |\Omega_{k+2}| = \frac{a_{k+2}}{q_{k+2}q_k} = \frac{a_{k+2}}{(a_{k+2}q_{k+1} + q_k)q_k} \geq \frac{1}{(q_{k+1} + q_k)q_k} . \quad (1.13a)$$

Taking this a stage further:

$$\frac{1}{(q_{k+1} + q_k)q_k} = \frac{1}{(a_{k+1}q_k + q_{k-1} + q_k)q_k} > \frac{1}{(a_{k+1} + 2)q_k^2} . \quad (1.13b)$$

We therefore arrive at the pleasantly similar forms for the close lower and upper bounds

$$\frac{1}{(a_{k+1} + 2)q_k^2} < \epsilon_k < \frac{1}{a_{k+1}q_k^2}. \quad (1.14)$$

These bounds are illustrated in Table 3 for the example used in Tables 1 and 2.

1.8 A graphical illustration of Eq 1.11: $\epsilon < 1/(2q^2)$

The relation $\epsilon < 1/(2q^2)$, giving a bound on error in terms of the denominator of a convergent, will be prevalent throughout this paper. There is a pleasing geometrical illustration of the paired relations in Eq 1.11, shown in Figure 3.

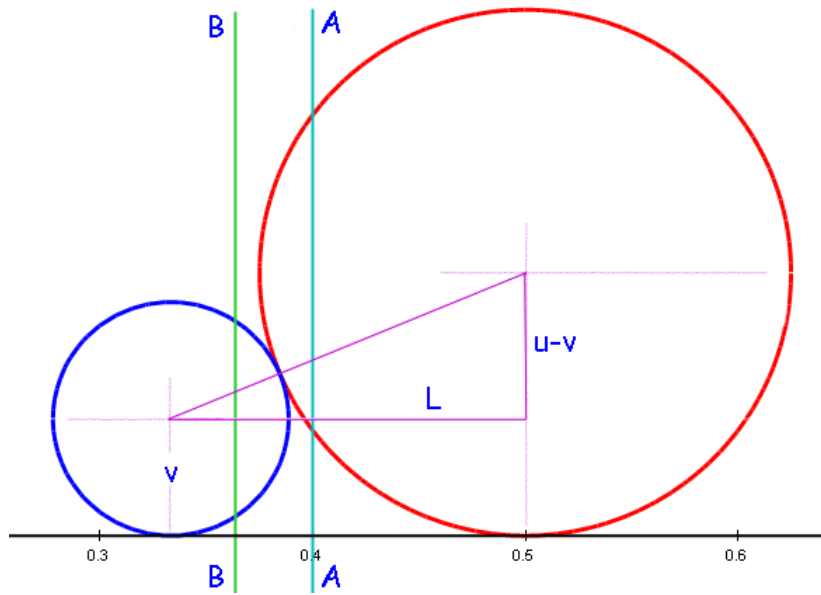


Figure 3: Circles associated with two adjacent convergents.

To each fraction p/q associate a circle on graph paper so that it is tangent to the x -axis and has x co-ordinate p/q . Now investigate the condition for two such circles, associated with adjacent convergents p/q and r/s , to be tangent to each other. If the radius of the p/q circle is u , and that of the r/s circle is v , the right-angled triangle in the figure satisfies Pythagoras' theorem:

$$L^2 + (u - v)^2 = (u + v)^2 \quad \text{so} \quad L^2 = 4uv.$$

$$\text{But } L = \frac{p}{q} - \frac{r}{s} = \frac{ps - qr}{qs} = \frac{1}{qs}$$

since from Eq 1.5 $ps - qr = 1$. The condition for the circles to be tangent is therefore that

$$u = \frac{1}{2q^2}, \quad v = \frac{1}{2s^2}.$$

Any real number θ to which p/q and r/s are convergents must lie between p/q and r/s . Two examples are the sea-green line marked A and the bright green line B in Figure 3. θ at A cuts the larger circle but misses the smaller one. Therefore the error in convergent p/q is less than $1/(2q^2)$, but error in r/s exceeds $1/(2s^2)$. On the other hand θ at B lies within one radius v of the convergent r/s ,

but more than u from p/q . A third case is those reals in a narrow band close to midway between p/q and r/s , for which θ lies inside both circles. For these both convergents have $\epsilon < 1/(2 \textit{denominator}^2)$. All this is contained in Eq 1.11.

That concludes this introductory overview of essential features. Later sections develop particular aspects in some detail.

2 Euclid's GCD algorithm

2.1 Expressing a rational number as a simple continued fraction

In §1.1 we had the example

$$\frac{305}{131} = \frac{262 + 43}{131} = 2 + \frac{43}{131} = 2 + \frac{1}{\frac{131}{43}}.$$

The decomposition of a rational number as a continued fraction uses Euclid's famous algorithm for the greater common divisor of numerator and denominator. This is the Euclidean decomposition:

$$305 = 2 \times 131 + 43$$

$$131 = 3 \times 43 + 2$$

$$43 = 21 \times 2 + \mathbf{1}$$

$$2 = 2 \times \mathbf{1} + 0$$

The continued fraction is given by the successive integer quotients as $\{2 : 3, 21, 2\}$. The $\gcd(305, 131) = 1$ (printed in bold), this being the final divisor which divides without remainder, and hence also the last non-zero remainder.

The reciprocal fraction $131/305$ has the same digits in its continued fraction as $305/131$, except they are moved along one place. To see why, consider that the Euclidean algorithm for $131/305$ begins

$$131 = 0 \times 305 + 131$$

then continues

$$305 = 2 \times 131 + 43 \text{ etc.}$$

just like the beginning of $305/131$. Hence $13/305$ is $\{0 : 2, 3, 21, 2\}$ and in general

$$\{0 : a, b, c, d, \dots\} = \frac{1}{\{a : b, c, d, \dots\}}. \quad (2.1)$$

It is also worth noting that all equivalent ordinary fractions have the same simple continued fraction. (Remember that in a simple continued fraction all the numerators, except for the integer part, are 1.) The continued fraction evaluates to an ordinary fraction in lowest terms; all factors common to numerator and denominator are cancelled, so $\gcd(\text{numerator}, \text{denominator})=1$. As an illustration, consider $36/356$:

$$356 = 9 \times 36 + 32$$

$$36 = 1 \times 32 + 4$$

$$32 = 8 \times 4 + 0$$

The continued fraction is $\{0 : 9, 1, 8\}$ and $\gcd(36, 356) = 4$. If we now evaluate successive convergents, we get $1/9$, $1/10$ and $9/89$, which is $36/356$ in lowest terms.

Euclid's gcd algorithm is proved in textbooks on number theory, but I give here my own account. Recall that if A and B are integers, $A > B$, there is a unique quotient q and remainder R , with R strictly less than B , such that

$$A = Bq + R, \quad 0 \leq R < B.$$

The key point is that $\gcd(A, B)$ equals $\gcd(B, R)$ — I give a proof below. So if we start with $A_1 = B_1q_1 + R_1$ then let B_1 be a new A_2 and R_1 a new B_2 , we can write $A_2 = B_2q_2 + R_2$ knowing that $\gcd(A_2, B_2) = \gcd(A_1, B_1)$. Eventually this step by step reduction must stop at some level n at which R_n exactly divides B_n . R_n (which might be 1) must therefore be $\gcd(B_n, R_n)$, and this is also $\gcd(A_1, B_1)$.

It remains to prove that $\gcd(A, B) = \gcd(B, R)$. Suppose that $\gcd(A, B) = g > 1$, so $A = ag$, $B = bg$ for some integers a, b . Then $A - Bq = g(a - bq) = R$. Since g divides the left side, it must divide the right, so $R = rg$ for some r . We can now divide through by g to get $a = bq + r$ where $\gcd(a, b) = 1$. It remains to prove that now b and r also have no common factor except 1. Just apply the same argument a second time. Suppose for the sake of argument that they do have a common factor $f \neq 1$ so that $b = \beta f$ and $r = \rho f$. Then $a = \beta f q + \rho f$ for some integers β and ρ . But since f divides the right side of this equation, it must also divide the left, implying that $a = \alpha f$ for some integer α . But this contradicts a and b having only 1 as a common factor.

2.2 Application of GCD to linear congruence equations

A frequent type of exam question in introductory number theory courses is

‘Find the general solution of the following Diophantine equations ²:

- i) $305x - 131y = 1$,
- ii) $1485x + 1745y = 15$.’

This type of linear equation can be solved using Euclid’s decomposition to express $305/131$ or $1485/1745$ respectively as a continued fraction. A key step is to use relation Eq 1.5, namely for adjacent convergents

$$p_k q_{k-1} - p_{k-1} q_k = (-1)^{k-1}.$$

Match this expression to our original equation in question i) by setting $p_k/q_k = 305/131$, and using the penultimate convergent $p_{k-1}/q_{k-1} = 149/64$, so

$$305 \times 64 - 131 \times 149 = 19520 - 19519 = 1.$$

This is only one solution to the equation; an infinity of others is found by adding and subtracting a constant K :

$$\begin{aligned} 305 \times 64 + K - 131 \times 149 - K &= 1. \\ 305 \left(64 + \frac{K}{305} \right) - 131 \left(149 + \frac{K}{131} \right) &= 1. \end{aligned}$$

To obtain a solution in integers we can take K to be any common multiple of 305 and 131: that is, any multiple of their least common multiple, 39955. So the general solution is

$$x = 64 + 131k, \quad y = 149 + 305k, \quad k \text{ any integer.}$$

The second equation ii) above is made only slightly more complicated by the fact that 1485 and 1745 are not coprime. Nevertheless, we proceed to express $1485/1745$ as a continued fraction and evaluate the penultimate convergent. The gcd is 5 and in lowest terms $1485/1745 = 297/349$.

²A Diophantine equation is one whose coefficients and solutions are all integers. The name comes from Diophantus of Alexandria who compiled a major book on arithmetic in the 3rd century AD.

The continued fraction is $(0 : 1, 5, 1, 2, 2, 7)$ and the penultimate convergent is $40/47$. We therefore obtain

$$297 \times 47 - 349 \times 40 = 13959 - 13960 = -1.$$

Turn this round, multiply by 3 ($=15/\text{gcd}$), add and subtract multiples of 297×349 , and multiply by the gcd:

$$\begin{aligned} 297 \times (-47) + 349 \times 40 &= 1. \\ 297 \times (-47) \times 3 + 349 \times 40 \times 3 &= 3. \\ 297 \left(-141 + \frac{297 \times 349k}{297} \right) + 349 \left(120 - \frac{297 \times 349k}{349} \right) &= 3, \quad k \text{ any integer.} \\ 1485(-141 + 349k) + 1745(120 - 297k) &= 15. \end{aligned}$$

We arrive at the general solution

$$x = -141 + 349k, \quad y = 120 - 297k, \quad k \text{ any integer.}$$

Note that there can be no solution unless $\text{gcd}(1485, 1745)$ divides the constant (15 in this case) on the right hand side.

2.3 Gradient and cutting sequence representations

This short section presents two closely related geometric representations of a real number r as a continued fraction. Another representation is given in §9.3.

We need first to introduce a two-dimensional ‘lattice’. This is a planar grid of points and their joining lines formed by translation of a unit cell, spanned by two basis vectors. A square lattice is the set of integer points (m, n) in the familiar x, y Cartesian co-ordinate frame, but in general the basis vectors will be neither perpendicular nor the same length.

Since the function $\tan^{-1} r$ is defined for all real r , r can be represented on a 2-D graph by a straight line \mathbf{L} with gradient $r = \tan \phi$. I’ll take the simple example of $r = 11/4 = 2.75$. Figure 4a shows this line, in red, against a square lattice, and 4b against a lattice in which the x -axis has been rotated until its gradient is 2. The basis vectors are in dark green. (Ignore the letters h and v for the moment.)

The essence of the continued fraction algorithm is iteratively to take out the integer part of r and invert the remainder. Graphically ‘taking out the integer part’ can be likened to reducing the gradient of \mathbf{L} by an integer such that its reduced gradient lies between 0 (horizontal) and 1 (45°). In our example this is the red line in Figure 4c with gradient $r - 2 = 0.75$. Another way of looking at this is to measure \mathbf{L} against a co-ordinate lattice like Figure 4b which has been skewed such that the gradient of the x basis vector is $\lfloor r \rfloor$, the integer part of r relative to the original square lattice. Whether we think of \mathbf{L} being compressed in the vertical direction or the lattice skewed upwards, we arrive at the red line in Figure 4c which represents the residual gradient $r - \lfloor r \rfloor$. The next step of ‘inverting the remainder’ corresponds geometrically to reflecting this red line in $x = y$ to produce the green line with gradient $4/3$ in Figure 4c. This green line is then the start of the next iteration, shown in Figure 4d against a lattice with x -axis at gradient 1. The process ends if and when a derived line has integer gradient. Because $11/4$ has continued fraction $\{2 : 1, 3\}$, we see the partial quotient 2 in the skewed lattice of Figure 4b, and the 1 in Figure 4d. In this way the number r is represented graphically as a sequence of lines at integer gradients – in our example 2, 1, 3.

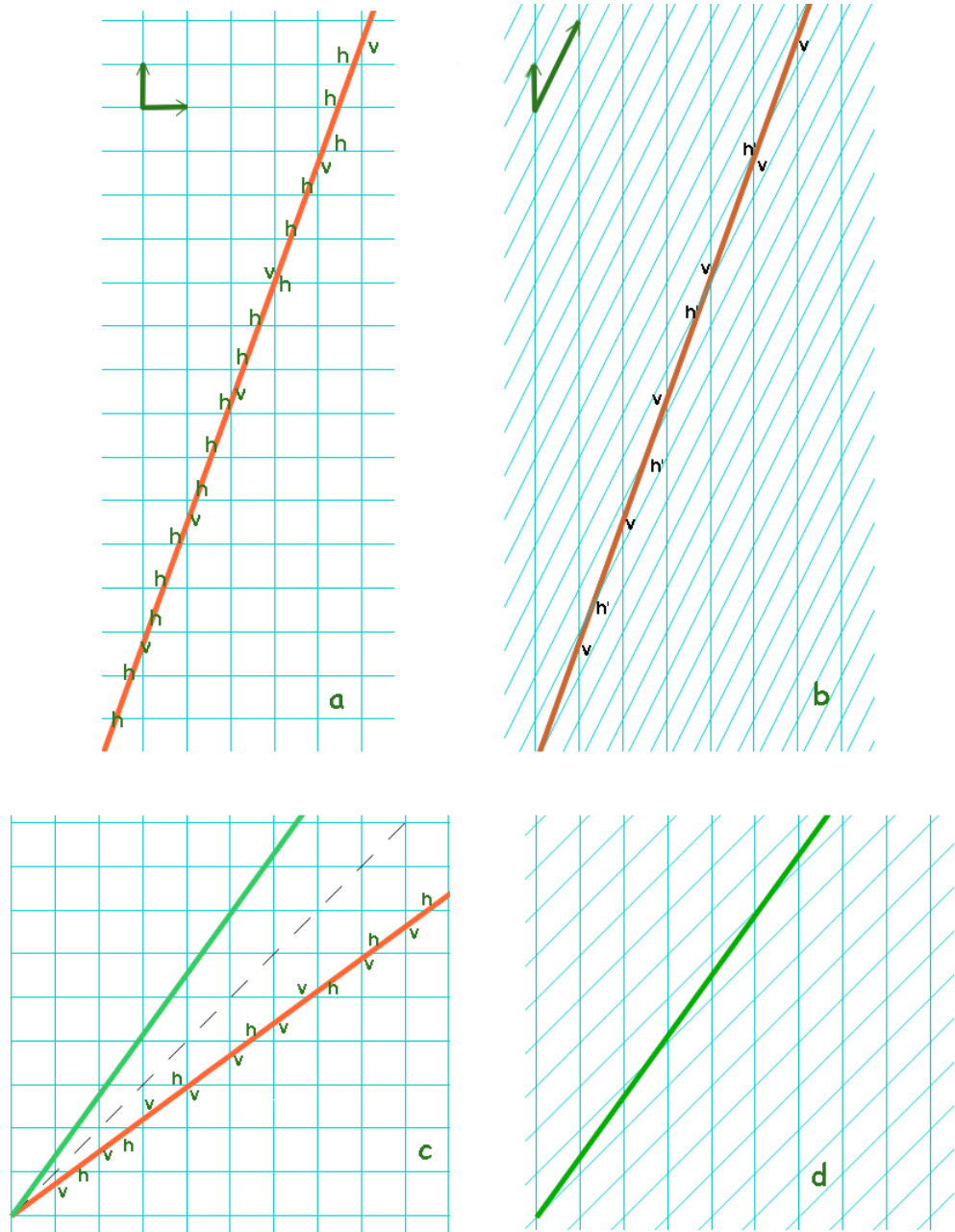


Figure 4: The red line \mathbf{L} has gradient $11/4$ and represents the real number $r = 2.75$. a) shows \mathbf{L} against a square lattice and b) against one with x -axis skewed to $\tan^{-1} 2$. c) and d) show the second iteration.

The second geometric representation of r is related to the first by describing the line \mathbf{L} and its reduced derivatives in terms of the sequence in which the line intersects the lattice grid lines. In Figure 4a imagine a tiny man walking along the red line in the upwards direction. As he crosses one of the pale blue grid lines, he notes whether that line is horizontal (h) or vertical (v). He records the sequence $\dots h v h h h v h v h v h h h v h h h v \dots = \dots h^2 v h^3 v h^2 v h^3 v h^2 v h^3 v \dots$, continuing indefinitely. Since the gradient exceeds 1, the h grid lines are cut more frequently than the v ones, and the letters v appear singly, separated by either two or three h . This sequence is called a ‘cutting sequence’ and, when taken to infinity, is characteristic of \mathbf{L} and hence the number r .

When \mathbf{L} is viewed against the highly skewed lattice with gradient 2 in Figure 4b, the cutting sequence is clearly different. Intersections with the vertical lattice lines are unchanged, but there are fewer intersections with the rotated x -axis lines, marked h' . The new cutting sequence is $\dots vh'vh'vh'vvh'v\dots$. Now look at the cutting sequence of the red line in Figure 4c: $vhvvhvvhvvh\dots$. Ignoring the dashes, these are the same, as we would expect because of the equivalence of rotating the lattice (4b) and vertically compressing the line (4c). This vertical compression is equivalent to removing some of the h from the cutting sequence of Figure 4a. In 4a the least number of h between any two v is 2. The compressed sequence is obtained by replacing each occurrence of vhh by v to get $\dots vhhvhhvhh\dots$. (Ignore the leading hh because they could be the end section of a vhh string.) The string reduction algorithm is therefore as follows: if $r > 1$ there is a preponderance of h and the v are isolated. Between each nearest pair of v will be either $[r]$ or $[r] + 1$ ' h '. Form the string $vh\dots h$ with $[r]$ ' h ' and replace its every occurrence by a single ' v '. The resulting string will have a preponderance of v , so repeat the procedure with h and v interchanged.

§8 in Part III will describe the use of matrices in respect of continued fractions. Anticipating this, observe that the string substitution $vh^{[r]} \rightarrow v$ is equivalent to the rotation of the x -axis in Figure 5 by $\tan^{-1}[r]$, and the latter is equivalent to a change of basis vector by the transformation matrix

$$\begin{pmatrix} 1 & 0 \\ [r] & 1 \end{pmatrix}.$$

In particular the base vectors $h = (1 \ 0)$ and $v = (0 \ 1)$ become respectively

$$\begin{pmatrix} 1 & 0 \\ [r] & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ [r] \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ [r] & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

In closing, note that in each case the line \mathbf{L} has been reduced to a line between 0° and 45° by subtracting from its gradient the largest integer less than r . In many cases $[r]$ will not be the closest integer to r . However, subtracting a number greater than r would lead to a different type of continued fraction, one having $-$ signs in the partial quotients.

2.4 The ambiguous final partial quotient, a_F

Now for some cautionary advice on evaluating convergents where the final partial quotient $a_F = 1$. A finite continued fraction such as $\{1: 2, 3, 4\}$ can also be written as $\{1: 2, 3, 3, 1\}$ since $3 + \frac{1}{1} = 4$. Both evaluate to $\frac{43}{30}$. However the extra terminal partial quotient generates an extra penultimate convergent:

$$\begin{aligned} \{1: 2, 3, 4\} & \text{ has convergents } \frac{1}{1}, \frac{3}{2}, \frac{10}{7}, \frac{43}{30} \\ \{1: 2, 3, 3, 1\} & \text{ has convergents } \frac{1}{1}, \frac{3}{2}, \frac{10}{7}, \frac{33}{23}, \frac{43}{30}. \end{aligned}$$

So what is the status of this extra convergent, $33/23$?

To answer this, we need a clear view of what it means to evaluate a continued fraction. Euclid's algorithm tells us to stop the process immediately the remainder is zero. On these grounds the authentic representation of $43/30$ is $\{1: 2, 3, 4\}$ because it has the Euclidean decomposition

$$43 = 1 \times 30 + 13, \quad 30 = 2 \times 13 + 4, \quad 13 = 3 \times 4 + 1, \quad 4 = 4 \times 1 + 0.$$

The authentic penultimate convergent of $43/30$ is $10/7$, obtained by truncating the continued fraction at $\{1: 2, 3\}$, which is equivalent to forcing the remainder at the previous iteration to be 0 (*i.e.* $13 \rightarrow 3 \times 4 + 0$).

The fraction $33/23$ is the value of $\{1 : 2, 3, 3\}$, obtained by truncating the 1 from $\{1 : 2, 3, 3, 1\}$. $33/23$ is *not* a convergent of $43/30$. Nevertheless, as explained in §7.4, Part III, it has some properties of a convergent, so I will term it a ‘quasi-convergent’. However, both $33/23$ and $43/30$ can be adjacent authentic convergents of a longer continued fraction, corresponding to a common fraction with larger denominator. For example

$$\{1 : 2, 3, 3, 1, 5\} \text{ has convergents } \frac{1}{1}, \frac{3}{2}, \frac{10}{7}, \frac{33}{23}, \frac{43}{30}, \frac{248}{173}.$$

Here $\frac{33}{23}$ is an authentic convergent. Indeed, in evaluating the sequence C_k of $\{1 : 2, 3, 3, 1, 5\}$, having deleted the final 5, it would be perverse to combine the penultimate 1 with the previous 3 – that would give a sequence $\frac{1}{1}, \frac{3}{2}, \frac{10}{7}, *, \frac{43}{30}, \frac{248}{173}$, with the authentic convergent $\frac{33}{23}$ omitted from position *.

2.5 Continued fractions with reversed a_k

This is a suitable juncture to describe the interesting relation between the finite fraction $\{a_0 : a_1, a_2, a_3, \dots, a_F\}$ and the fraction whose a_k are in reverse order, $\{0 : a_F, a_{F-1}, \dots, a_1\}$ (with a_0 omitted). Recall the somewhat similar property described in §2.1, Eq 2.1 where shifting the partial quotients by one place is equivalent to taking the reciprocal of the number.

A numerical example will help. Take $\theta = \{4 : 3, 5, 2, 7, 6\}$. The convergents are

$$\frac{4}{1}, \frac{13}{3}, \frac{69}{16}, \frac{151}{35}, \frac{1126}{261}, \frac{6907}{1601}.$$

Now form fractions q_{k-1}/q_k from adjacent denominators:

$$\frac{261}{1601} = \{0 : 6, 7, 2, 5, 3\}, \quad \frac{35}{261} = \{0 : 7, 2, 5, 3\}, \quad \frac{16}{35} = \{0 : 2, 5, 3\}, \quad \text{etc.}$$

This property is readily shown from the recursion relations Eq 1.2. If a_F is the final partial quotient of θ ,

$$\frac{q_{F-1}}{q_F} = \frac{q_{F-1}}{a_F q_{F-1} + q_{F-2}} = \frac{1}{a_F + \frac{q_{F-2}}{q_{F-1}}}, \quad (2.2)$$

so the leading partial quotient is a_F . Continuing, the next quotient will be a_{F-1} , and so on.

There is a small complication when the given continued fraction for θ starts with 1 as in $\{0 : 1, 2, 3, 4, 5, 6\}$. From §1.2, the successive convergents of $\{0 : 1, 2, 3, 4, 5, 6\}$ are

$$\frac{1}{1}, \frac{2}{3}, \frac{7}{10}, \frac{30}{43}, \frac{157}{225}, \frac{972}{1393},$$

whilst the convergents of $\{0 : 6, 5, 4, 3, 2, 1\}$ are

$$\frac{1}{6}, \frac{5}{31}, \frac{21}{130}, \frac{68}{421}, \frac{157}{972}, \frac{225}{1393}.$$

A straight Euclidean decomposition of $q_{F-1}/q_F = 225/1393$ gives $\{0 : 6, 5, 4, 3, 3\}$, but this can be converted immediately into the required form $\{0 : 6, 5, 4, 3, 2, 1\}$ by using $3 = 2 + \frac{1}{1}$. Note also that the truncation $\{0 : 6, 5, 4, 3, 2\} = 157/972 = p_{F-1}/p_F$.

Eq 2.2 can be combined with Eqs 1.3, 1.8 to give an alternative exact expression for the error ϵ_k in the k^{th} convergent of some given number θ :

$$\epsilon_k = |C_k - \theta| = \frac{1}{q_k(\theta_{k+1}q_k + q_{k-1})} = \frac{1}{q_k^2(\theta_{k+1} + \chi_k)}, \quad \theta_{k+1} > 1, \quad \chi_k = \frac{q_{k-1}}{q_k}$$

from which

$$\epsilon_k q_k^2 = \frac{1}{\theta_{k+1} + \{0 : a_k, a_{k-1}, \dots, a_2, a_1\}} = \frac{1}{\{a_{k+1} : a_{k+2}, \dots, \} + \{0 : a_k, a_{k-1}, \dots, a_2, a_1\}} \quad (2.3)$$

$$\epsilon_k q_k^2 = \frac{1}{\theta_{k+1} + \chi_k}.$$

This expresses the error relative to q_k^2 as a sum of two continued fractions, both derived by splitting the continued fraction for θ at position k . This form of expressing the error will be referred to in §10.4 and §11.4. We shall show in §3 that in the special case where θ is a quadratic surd with a recurring sequence of partial quotients, the two halves of the split continued fraction are such that the one is the negative algebraic conjugate of the other.

3 Recurring continued fractions and quadratic surds

One of the intriguing properties of continued fractions is that *any* infinitely recurring continued fraction represents a solution of a quadratic equation. For example $\sqrt{2} = \{1 : \underline{2}\}$, where the underlining denotes a recurring sequence of partial quotients. You might care to reflect on the parallel properties of recurring decimals to represent rational numbers. *Any* rational number can be represented as a recurring decimal. Where the numbers have only 2 and 5 as prime factors, their decimal terminates because the recurring digit is 0. Any non-terminating, non-recurring decimal is irrational — either an algebraic or a transcendental number.

For completeness, here is a brief look at the finite continued fraction with single recurring a_k . Then the section analyses the more interesting and important infinite recurring continued fractions, from two points of view:

1. given a recurring continued fraction, to determine the quadratic surd which it represents. This will have the form $\alpha + \beta\sqrt{\gamma}$ (all integers)
2. to describe an algorithm for finding the continued fraction representation of a chosen integer square root, \sqrt{N} .

In §7, Part III, in the context of symmetry groups associated with continued fractions, there is some analysis of the transition from a finite continued fraction having a recurring sequence to its infinitely recurring counterpart.

3.1 A finite recurring continued fraction

The simplest case of a recurring continued fraction has a single recurring partial quotient $\{a : a, a, \dots, a\}$. Repeated application of the recursion relation Eq 1.2 (or using matrix multiplication as Eq 1.4) gives the following polynomials in a :

$$\begin{aligned}
 p_0 &= q_1 = a \\
 p_1 &= q_2 = a^2 + 1 \\
 p_2 &= q_3 = a^3 + 2a \\
 p_3 &= q_4 = a^4 + 3a^2 + 1 \\
 p_4 &= q_5 = a^5 + 4a^3 + 3a \\
 p_5 &= q_6 = a^6 + 5a^4 + 6a^2 + 1 \\
 p_6 &= q_7 = a^7 + 6a^5 + 10a^3 + 4a \\
 p_7 &= q_8 = a^8 + 7a^6 + 15a^4 + 10a^2 + 1 \\
 p_8 &= q_9 = a^9 + 8a^7 + 21a^5 + 20a^3 + 5a \\
 p_9 &= q_{10} = a^{10} + 9a^8 + 28a^6 + 35a^4 + 15a^2 + 1 \\
 p_{10} &= q_{11} = a^{11} + 10a^9 + 36a^7 + 56a^5 + 35a^3 + 6a, \quad \text{etc.}
 \end{aligned}$$

Viewed diagonally through this array, the coefficients can be recognised as binomial coefficients ${}^n C_r$. The general formula for p_{k-1}, q_k is

$$p_{k-1} = q_k = \sum_{j=0}^{\lfloor \frac{k+1}{2} \rfloor} \binom{k-j}{j} C_j a^{(k-2j)} \tag{3.1}$$

where the floor function $\lfloor x \rfloor$ denotes the integer part of x . Note, in passing, that for $a = 1$ the series of polynomials evaluate to 1, 2, 3, 5, 8, 13, ..., the celebrated Fibonacci series.

I do not wish to spend more time on this finite case lest it be a distraction from the more important infinite case $\{a : a, a, \dots\}$. However some comment is appropriate on the transition to the infinite case. Clearly the convergent C_k tends to a as $a \rightarrow \infty$. Higher terms in this approximation can be determined by expanding C_k for some large k as a Taylor series in $1/a$ about $1/a = 0$, :

$$C_k = \frac{p_k}{q_k} \rightarrow a + \frac{1}{a} - \frac{1}{a^3} + \frac{2}{a^5} - \frac{5}{a^7} + \frac{14}{a^9} - \frac{42}{a^{11}} + \dots \quad (3.2)$$

It happens that this is a useful numerical approximation even for a as small as 3 – the value is $3 \cdot 302776$. The coefficients in this Taylor expansion are the Catalan numbers 1, 2, 5, 14, 42, 132, 429, 1430, 4862, ... = ${}^{2n}C_n/(n+1)$.

3.2 Infinite recurring continued fractions are quadratic surds

In this article the length of the recursion sequence is denoted by L . We consider explicitly the cases $L = 1, L = 2$.

3.2.1 Single recurring partial quotient

Consider the simplest infinite case where $\theta = \{a : \underline{a}\}$:

$$\theta = a + \frac{1}{a + \frac{1}{a + \frac{1}{\dots}}}$$

This can be arranged as the quadratic equation

$$\theta^2 - a\theta - 1 = 0$$

with solutions

$$\{a : \underline{a}\} = \frac{1}{2}(a + \sqrt{a^2 + 4}), \quad \{0 : \underline{a}\} = \frac{1}{2}(a - \sqrt{a^2 + 4}). \quad (3.3)$$

since the two roots are reciprocals of each other. The first formula can be used to evaluate square roots of the form $\sqrt{\text{integer}^2 + 4}$, such as the roots of 5, 8, 13, etc. Indeed, putting $a = 3$ gives $\frac{1}{2}(3 + \sqrt{13}) = 3 \cdot 302776$ in agreement with the limiting value found above, Eq 3.2. A widely recognised case is $\{1 : \underline{1}\} = \frac{1}{2}(1 + \sqrt{5}) \approx 1 \cdot 6180$, the Golden Ratio G of classical art and architecture. The partial convergents of this continued fraction are $\frac{3}{2}, \frac{5}{3}, \frac{8}{5}, \frac{13}{8}$, etc. formed from the Fibonacci series, as noted in §3.1. Its reciprocal is $1/G = \{0 : \underline{1}\} = G - 1$. Another special case is the fancifully called ‘silver ratio’ $\{2 : \underline{2}\} = 1 + \sqrt{2}$.

Where the recurring partial quotient a is even, the factor of $\frac{1}{2}$ in θ cancels and the continued fraction has a particularly simple form $\{a : \underline{2a}\} = \sqrt{a^2 + 1}$. For the record

$$\begin{aligned} a = 2 &\rightarrow \sqrt{8} = 2\{0 : \underline{2}\} + 2 \rightarrow \sqrt{2} = \{1 : \underline{2}\} \\ a = 4 &\rightarrow \sqrt{20} = 2\{0 : \underline{4}\} + 4 \rightarrow \sqrt{5} = \{2 : \underline{4}\} \\ a = 6 &\rightarrow \sqrt{40} = 2\{0 : \underline{6}\} + 6 \rightarrow \sqrt{10} = \{3 : \underline{6}\} \\ a = 8 &\rightarrow \sqrt{68} = 2\{0 : \underline{8}\} + 8 \rightarrow \sqrt{17} = \{4 : \underline{8}\} \end{aligned}$$

$$\begin{aligned}
a = 10 &\rightarrow \sqrt{104} = 2\{0 : \underline{10}\} + 10 \rightarrow \sqrt{26} = \{5 : \underline{10}\} \\
a = 12 &\rightarrow \sqrt{148} = 2\{0 : \underline{12}\} + 12 \rightarrow \sqrt{37} = \{6 : \underline{12}\} \\
a = 14 &\rightarrow \sqrt{200} = 2\{0 : \underline{14}\} + 14 \rightarrow \sqrt{50} = \{7 : \underline{14}\}.
\end{aligned}$$

Since $\sqrt{50} = 10/\sqrt{2}$, this last relation allows us to invent ‘factorisations’ in continued fractions such as $\{1 : \underline{2}\} \times \{7 : \underline{14}\} = 10$.

3.2.2 Double recurring partial quotients

The extension to two recurring partial quotients $\{a : \underline{b}, a\}$ is readily made. Let

$$\theta = a + \frac{1}{b + \frac{1}{a + \frac{1}{b + \frac{1}{a + \dots}}}} = a + \frac{1}{b + \frac{1}{\theta}}$$

so

$$b\theta^2 - ab\theta - a = 0.$$

The required solution for $a \neq 0$ has the positive root:

$$\theta = \theta_+ = \{a : \underline{b}, a\} = \frac{1}{2b} [ab + \sqrt{a^2b^2 + 4ab}] > 1. \quad (3.4)$$

The reciprocal of this is < 1 and, from Eq 2.1, has the partial quotients shifted one place to the right. This, therefore, is related to Eq 3.4 through interchange of a and b :

$$b + \frac{1}{\theta_+} = \{b : \underline{a}, b\} = \frac{1}{2a} [ab + \sqrt{a^2b^2 + 4ab}] > 1.$$

Since the terms in brackets $[\dots]$ are symmetrical in a and b ,

$$b\{a : \underline{b}, a\} = a\{b : \underline{a}, b\}. \quad (3.5)$$

The other root, θ_- , with the negative square root, has negative value. Denoting $a^2b^2 + 4ab$ by D , the discriminant, its reciprocal is

$$\frac{1}{\theta_-} = \frac{2b(ab + \sqrt{D})}{(ab - \sqrt{D})(ab + \sqrt{D})} = -\frac{ab + \sqrt{D}}{2a} = \{b : \underline{a}, b\}.$$

This is an important observation: with quadratic surds, the conjugate roots are related through the partial quotients being in reverse order. Recall §2.4.

3.2.3 Multiple recurring partial quotients

The cases of three recurring partial quotients are readily calculated, if rather unwieldy. For many integers N they allow \sqrt{N} to be expressed in several forms involving multiples of recurring continued fractions. Examples include $\sqrt{37} = \{6 : \underline{12}\} = 4+3\{0 : \underline{1}, \underline{2}, \underline{3}\} = 5+3\{0 : \underline{2}, \underline{1}, \underline{3}\} = 5+4\{0 : \underline{3}, \underline{1}, \underline{2}\}$.

We have come close to seeing why any recurring continued fraction must evaluate to a surd involving only a square root. The general case is

$$\theta = \{a : \underline{b}, \underline{c}, \dots, \underline{y}, \underline{z}, a\} = a + \frac{1}{b + \frac{1}{c + \frac{1}{\dots + \frac{1}{z + \frac{1}{\theta}}}}}$$

No matter the length of the recurring sequence, there are only two θ here, and in the evaluation of the continued fraction they become multiplied to give a quadratic equation.

I also emphasise the fact that quadratic surds which are algebraic conjugates of each other have continued fractions which are related through the partial quotients being in reverse order. Thus if θ above satisfies the quadratic $A\theta^2 + B\theta + C = 0$ with roots $\theta_+ > 0$, $\theta_- < 0$, then $\chi = \{0 : \underline{z, y, x, \dots b, a}\}$ satisfies $A\chi^2 - B\chi + C = 0$ with roots χ_+ , χ_- . Therefore

$$\theta_+ = -\chi_- > 1, \quad \theta_- = -\chi_+ > -1.$$

Quadratic surds also result if the recurring part of the continued fraction is a ‘tail’ following a finite number of non-recurring partial quotients. The infinite $\{a : b, c, \dots, n, \underline{r, s, t, \dots, z}\}$ can be evaluated as the finite continued fraction $\{a : b, c, \dots, n, \xi\}$, where ξ is a quadratic surd satisfying the equation $\xi = \{r : s, t, \dots, z, \xi\}$. An analogous situation exists with decimal representations of fractions such as $\frac{5}{24} = 0.208\bar{3}$, where the recurring part follows several non-recurring digits.

3.3 Expressing \sqrt{N} as a continued fraction

We now look at quadratic surds from another angle by asking how to find the continued fraction corresponding to the square root of a given integer, N .

The algorithm essentially involves repeatedly extracting the integer part from an improper fraction (numerator $>$ denominator) and is usually presented in textbooks as follows. Take the case of $\sqrt{103}$, which has quite a long recursion sequence. Since $10^2 < 103 < 11^2$, deduct the integer part and write

$$\sqrt{103} = 10 + (\sqrt{103} - 10) = 10 + \frac{1}{\frac{1}{\sqrt{103} - 10}}.$$

Either use a calculator to find the reciprocal of the remainder, or rationalise the denominator by determining α and β such that

$$\frac{1}{\sqrt{103} - 10} = \alpha\sqrt{103} + \beta.$$

Multiply, equate rational and surd components, and so find that $\alpha = 1/3$, $\beta = 10/3$. Hence

$$\sqrt{103} = 10 + (\sqrt{103} - 10) = \frac{20}{3} + \frac{\sqrt{103} - 10}{3}$$

and the integer part of this is 6. The continued fraction therefore begins $\{10 : 6, \dots\}$. The next step is to find the reciprocal of $(\sqrt{103} - 8)/3$.

This method is mathematically sound and is easy to carry out by hand with a calculator. Nevertheless, I find it more transparent and satisfactory to follow the procedure below, which involves first writing down a ‘look-up table’ of quadratic factors for \sqrt{N} from which the continued fraction can be quickly read. Let $N = n^2 + k$ where $n = \lfloor \sqrt{N} \rfloor$ is the largest integer less than \sqrt{N} . Also let $\delta < 1$ be the fractional part of \sqrt{N} . Thus

$$\sqrt{N} = n + \delta = n + \frac{1}{\frac{1}{\delta}},$$

$$N = n^2 + k = (n + \delta)^2 \quad \text{so} \quad k = 2n\delta + \delta^2 = (2n + \delta)(0 + \delta).$$

Table 4: List of paired quadratic factors of integers $N - j^2$

j	quadratic	value
0	$(n + 0 + \delta)(n - 0 + \delta)$	$N = n^2 + k$
1	$(n + 1 + \delta)(n - 1 + \delta)$	$N - 1 = n^2 + k - 1$
2	$(n + 2 + \delta)(n - 2 + \delta)$	$N - 4 = n^2 + k - 4$
3	$(n + 3 + \delta)(n - 3 + \delta)$	$N - 9 = n^2 + k - 9$
...
$n - 2$	$(2n - 2 + \delta)(2 + \delta)$	$N - (n - 2)^2 = k + 4n - 4$
$n - 1$	$(2n - 1 + \delta)(1 + \delta)$	$N - (n - 1)^2 = k + 2n - 1$
n	$(2n + \delta)(0 + \delta)$	$N - n^2 = k$

This factorisation is rearranged to give the required first reciprocal:

$$\frac{1}{\delta} = \frac{2n + \delta}{k}.$$

We will also need algebraic factorisations of $N - j^2$, j an integer with $0 \leq j \leq n$. These have the form

$$n^2 - j^2 + k = (n + \delta)^2 - j^2 = (n + j + \delta)(n - j + \delta).$$

A table like Table 4 can be quickly written down for these. Using this table one can readily follow the chain of repeated factorisation, inversion and finding remainder through values of j until it eventually and inevitably returns to the quadratic factor $0 + \delta$, which was the starting value. From this point the chain repeats itself round an endless loop, so at this stage we have the recurring continued fraction for \sqrt{N} .

The example of $\sqrt{103}$ should make the method clear. The required table of factorisations is Table 5. A typical step in the procedure would be to use the quadratic relation

$$(18 + \delta)(2 + \delta) = 103 - 8^2 = 39 \quad \text{so} \quad \frac{3}{2 + \delta} = \frac{18 + \delta}{3 \times 13} = 1 + \frac{5 + \delta}{13},$$

then take the reciprocal of $(5 + \delta)/13$ as the next step. The cancellation of the factor 3 between numerator and denominator is crucial, and similar cancellation occurs at each subsequent step. This is explained in §3.4

In finding $\sqrt{103}$ we start at the bottom row of Table 5 and read off the following sequence of fractions. The order of using the factorisations is given in column 3. Note how this sequence doubles back on itself after the half-way position.

$$\begin{aligned} \frac{1}{\delta} &= \frac{20 + \delta}{3} = 6 + \frac{2 + \delta}{3} \\ \frac{3}{2 + \delta} &= \frac{18 + \delta}{13} = 1 + \frac{5 + \delta}{13} \\ \frac{13}{5 + \delta} &= \frac{15 + \delta}{6} = 2 + \frac{3 + \delta}{6} \\ \frac{6}{3 + \delta} &= \frac{17 + \delta}{9} = 1 + \frac{8 + \delta}{9} \end{aligned}$$

Table 5: List of paired quadratic factors for evaluating $\sqrt{103}$

j	$n + j, n - j$	order	quadratic	value & integer factors
0	10, 10		$(10 + \delta)(10 + \delta)$	$= 103$
1	11, 9		$(11 + \delta)(9 + \delta)$	$= 103 - 1 = 102$
2	12, 8	5, 8	* $(12 + \delta)(8 + \delta)$	$= 103 - 4 = 99 = 9 \times 11$
3	13, 7		$(13 + \delta)(7 + \delta)$	$= 103 - 9 = 94$
4	14, 6		$(14 + \delta)(6 + \delta)$	$= 103 - 16 = 87$
5	15, 5	3, 10	* $(15 + \delta)(5 + \delta)$	$= 103 - 25 = 78 = 6 \times 13$
6	16, 4		$(16 + \delta)(4 + \delta)$	$= 103 - 36 = 67$
7	17, 3	4, 9	* $(17 + \delta)(3 + \delta)$	$= 103 - 49 = 54 = 9 \times 6$
8	18, 2	2, 11	* $(18 + \delta)(2 + \delta)$	$= 103 - 64 = 39 = 13 \times 3$
9	19, 1	6, 7	* $(19 + \delta)(1 + \delta)$	$= 103 - 81 = 22 = 11 \times 2$
10	20, 0	1, 12	* $(20 + \delta)(0 + \delta)$	$= 103 - 100 = 3 = 1 \times 3$

$$\frac{9}{8 + \delta} = \frac{12 + \delta}{11} = 1 + \frac{1 + \delta}{11}$$

$$\frac{11}{1 + \delta} = \frac{19 + \delta}{2} = 9 + \frac{1 + \delta}{2}$$

$$\frac{2}{1 + \delta} = \frac{19 + \delta}{11} = 1 + \frac{8 + \delta}{11}$$

$$\frac{11}{8 + \delta} = \frac{12 + \delta}{9} = 1 + \frac{3 + \delta}{9}$$

$$\frac{9}{3 + \delta} = \frac{17 + \delta}{6} = 2 + \frac{5 + \delta}{6}$$

$$\frac{6}{5 + \delta} = \frac{15 + \delta}{13} = 1 + \frac{2 + \delta}{13}$$

$$\frac{13}{2 + \delta} = \frac{18 + \delta}{3} = 6 + \frac{\delta}{3}$$

$$\frac{3}{\delta} = 20 + \delta$$

We arrive back at remainder δ . Consequently the sequence 6, 1, 2, 1, 1, 9, 1, 1, 2, 1, 6, 20 recurs, being the integer parts of the successive fractions. We have found that $\sqrt{103}$ has the length 12 recursion sequence $\{10 : \underline{6, 1, 2, 1, 1, 9, 1, 1, 2, 1, 6, 20}\}$.

This method can readily be converted to a recursive algorithm and implemented on a computer in a few lines of code. It does not require a table like Table 5 actually to be created – all that is need is to keep track of j . Moreover, the method only requires integer division and modular functions (DIV and MOD). Therefore even very long recursion sequences can be calculated, limited only the largest integer which the computer can hold in memory. I have implemented this in a BASIC program and calculated, for example, $\sqrt{93452367}$ which has a recursion length $L = 528$. Of course, evaluating this as an ordinary fraction will be limited by the floating point precision of the computer, but calculating the sequence of partial quotients is not. This means that calculating the

continued fraction for \sqrt{N} can be regarded as always straightforward and without limitation, except for the integer-handling capacity of the computer when N is large.

3.4 Patterns in the partial quotients of \sqrt{N}

Note the palindromic form³ of the above sequence of partial quotients for $\sqrt{103}$. It has mirror symmetry about the central digit 9, with the exception that the last quotient is 20, not 10. This is a general result, as three other roots will illustrate:

$$\begin{aligned}\sqrt{31} &= \{5 : \underline{1, 1, 3, 5, 3, 1, 1, 10}\} \\ \sqrt{43} &= \{6 : \underline{1, 1, 3, 1, 5, 1, 3, 1, 1, 12}\} \\ \sqrt{46} &= \{6 : \underline{1, 3, 1, 1, 2, 6, 2, 1, 1, 3, 1, 12}\}\end{aligned}$$

The general pattern is

$$\sqrt{N} = \{a : \underline{b, c, d, e, \dots, e, d, c, b, 2a}\}. \quad (3.6)$$

Not all of the quadratic pairs in Table 5 have been used in finding $\sqrt{103}$. The six which are, marked * and ordered in column 3, follow from each other depending on how the integer value of $N - j^2$ in the last column of Table 5 factorises. There is an analogy with the placing of dominoes in that well known board game. Dominoes are played in an end-to-end ‘domino chain’ of paired integers. The domino chain here is $(1 \times 3) : (3 \times 13) : (13 \times 6) : (6 \times 9) : (9 \times 11) : (11 \times 2) : (2 \times 11) : (11 \times 9) : (9 \times 6) : (6 \times 13) : (13 \times 3) : (3 \times 1)$. Note the mirror symmetry about the central number 2. The only quadratic factorisations of Table 5 featuring inside the chain (not at the start, centre or end) are those for which *two* integers of the form $N - j^2$ share a common factor. But, if a factor, q say, occurs in only one integer $N - j^2$, the chain must reverse about q according to $\dots (p \times q) : (q \times p) \dots$. Therefore in all cases when the chain reaches a certain point (2 in the case of $\sqrt{103}$), its only option is to return along the path it came. This creates the palindrome.

To gain a deeper understanding of this domino chain, it is necessary to see why, at each step, factors always cancel between numerator and denominator. Here is an explanation for the first step; later steps are similar though more complicated to analyse algebraically. The first reciprocal is

$$\frac{1}{\delta} = \frac{2n + \delta}{k} = \lfloor \frac{2n}{k} \rfloor + \frac{n - j_1 + \delta}{k}$$

where $n - j_1 = 2n \bmod k$. At the next step the reciprocal is

$$\frac{k}{n - j_1 + \delta} = \frac{k(n + j_1 + \delta)}{N - j_1^2}.$$

The crucial property here is that k divides $N - j_1^2$. To see why

$$\begin{aligned}N - j_1^2 &= N - [n - (2n \bmod k)]^2 \\ &= n^2 + k - n^2 - (2n \bmod k)[2n - (2n \bmod k)]\end{aligned}$$

and $k|M - (M \bmod k)$ for all k, M . (Let $M = mk + \varepsilon$ for $\varepsilon < k$. Then $\varepsilon = M \bmod k$.)

³A palindrome is a word or sentence which has mirror symmetry in its letters. Examples in English include “Was it a rat I saw?” and the Napoleonic “Able was I ere I saw Elba”.

Incidentally, the doubling of n to $2n$ at the last partial quotient in the recursion sequence is a consequence of the $2n$ in $(2n + \delta)(0 + \delta) = N - n^2 = k$.

In his very readable book ‘The Higher Arithmetic’ (page 96 *et seq.*) Harold Davenport explains the palindromic symmetry in terms of the two roots, one with $+$, the other with $-$, of the quadratic for the recurring sequence (see §3.2). He shows that these have continued fractions whose recurring sequences are the reverse of each other.

I have noticed that the length L of recursion sequences is an even number far more often than it is odd. Some numerical work gives these statistics: between $N = 1000$ and $N = 1200$ 85% of sequences of \sqrt{N} are even, rising to 89.5% between 78,000 and 78200, and rising slightly again to 90.5% between 160,000 and 160,200.

3.4.1 Short recurring sequences in \sqrt{N}

For some values of k in $N = n^2 + k$, the ‘domino chain’ and hence the recursion sequence is quite short.

1. The shortest ones are for $k = 1$. The quadratic factorisation in the last row of Table 4 is

$$(2n + \delta)(0 + \delta) = N - n^2 = k = 1.$$

Since $\sqrt{N} = n + \delta$ and $1/\delta = 2n + \delta$, we arrive immediately back at the start. Examples of this $\{ n : \underline{2n} \}$ form are listed in §3.2.1 – for example, $\sqrt{5} = \{ 2 : \underline{4} \}$.

2. The case $k = 2$, $N = n^2 + 2$ has two recurring partial quotients. First

$$j = n : (2n + \delta)(0 + \delta) = N - n^2 = k = 2 \quad \text{so} \quad \frac{1}{\delta} = n + \frac{\delta}{2}.$$

The next required reciprocal is therefore $2/\delta = 2n + \delta$ and we are back to the start. Examples of this $\{ n : \underline{n}, \underline{2n} \}$ form are $\sqrt{3} = \{ 1 : \underline{1}, \underline{2} \}$, $\sqrt{6} = \{ 2 : \underline{2}, \underline{4} \}$, $\sqrt{11} = \{ 3 : \underline{3}, \underline{6} \}$ and $\sqrt{18} = \{ 4 : \underline{4}, \underline{8} \}$.

3. When $k = 3$, $N = n^2 + 3$, we have to determine the integer part of $1/\delta = (2n + \delta)/3$. There are three cases: $3 \mid n$, $3 \mid 2n - 1$ and $3 \mid 2n - 2$. As far as I can deduce, only the case $3 \mid n$ leads to a necessarily short recursion sequence:

$$\frac{1}{\delta} = \frac{2n}{3} + \frac{\delta}{3} \quad \text{so that} \quad \frac{3}{\delta} = 2n + \delta$$

with remainder δ takes us straight back to the start. Examples of this $\{ n : \underline{2n/3}, \underline{2n} \}$ form are $\sqrt{12} = \{ 3 : \underline{2}, \underline{6} \}$, $\sqrt{39} = \{ 6 : \underline{4}, \underline{12} \}$, $\sqrt{84} = \{ 9 : \underline{6}, \underline{18} \}$ and $\sqrt{147} = \{ 12 : \underline{8}, \underline{24} \}$.

4. The above analysis extends to all cases where k divides $2n$ because recursion requires only one quadratic factorisation. Specifically

$$k = (2n + \delta)\delta \quad \text{so} \quad \frac{1}{\delta} = \frac{2n}{k} + \frac{\delta}{k} \quad \text{and} \quad \frac{k}{\delta} = 2n + \delta.$$

Table 6 lists examples for various values of n and k .

k	\sqrt{N}	continued fraction	k	\sqrt{N}	continued fraction
4	$\sqrt{8}$	$\{ 2 : \underline{1}, 4 \}$	6	$\sqrt{87}$	$\{ 9 : \underline{3}, 18 \}$
4	$\sqrt{20}$	$\{ 4 : \underline{2}, 8 \}$	6	$\sqrt{231}$	$\{ 15 : \underline{5}, 30 \}$
4	$\sqrt{40}$	$\{ 6 : \underline{3}, 12 \}$	7	$\sqrt{56}$	$\{ 7 : \underline{2}, 14 \}$
4	$\sqrt{68}$	$\{ 8 : \underline{4}, 16 \}$	7	$\sqrt{203}$	$\{ 14 : \underline{4}, 28 \}$
4	$\sqrt{104}$	$\{ 10 : \underline{5}, 20 \}$	8	$\sqrt{24}$	$\{ 4 : \underline{1}, 8 \}$
4	$\sqrt{148}$	$\{ 12 : \underline{6}, 24 \}$	8	$\sqrt{72}$	$\{ 8 : \underline{2}, 16 \}$
5	$\sqrt{30}$	$\{ 5 : \underline{2}, 10 \}$	8	$\sqrt{152}$	$\{ 12 : \underline{3}, 24 \}$
5	$\sqrt{105}$	$\{ 10 : \underline{4}, 20 \}$	10	$\sqrt{35}$	$\{ 5 : \underline{1}, 10 \}$
5	$\sqrt{230}$	$\{ 15 : \underline{6}, 30 \}$	12	$\sqrt{48}$	$\{ 6 : \underline{1}, 12 \}$
6	$\sqrt{15}$	$\{ 3 : \underline{1}, 6 \}$	14	$\sqrt{63}$	$\{ 7 : \underline{1}, 14 \}$
6	$\sqrt{42}$	$\{ 6 : \underline{2}, 12 \}$	18	$\sqrt{99}$	$\{ 9 : \underline{1}, 18 \}$

Table 6: Some examples of \sqrt{N} where $N = n^2 + k$ and $k|2n$

3.4.2 Long recurring sequences in \sqrt{N}

In tables of factorisations such as Table 5 for $\sqrt{103}$ there are $n + 1$ rows. In principle the order of visiting the rows might take any course, subject to the domino chain and palindrome requirements described previously. On these grounds the length L of the recurring sequence might be expected to be as large as $2(n + 1)$ or even longer. However, some numerical investigation with fairly small integers leads me to think that few \sqrt{N} have sequences longer than $1 \cdot 5\sqrt{N}$. Table 7 gives some examples. By empirically curve fitting I find that the larger L are roughly proportional to \sqrt{N} . Typical long lengths are about $L = 1 \cdot 3\sqrt{N}$, while the upper envelope is about $L = 1 \cdot 8\sqrt{N}$. Four N close to this upper envelope are 46, 211, 214 and 2011 (year of writing!). N with large L for \sqrt{N} are usually separated by several integers with small and moderate L ; I have not observed two large- L consecutive integers. Some larger L occur for N just less than the next square, $(n + 1)^2$: *e.g.* 46, 94, 166, 214, 244. There are also quite a lot with the form $N = n^2 + k$ with $k = 3$ or a multiple of 3: *e.g.* 7, 19, 67, 103, 124. However, I see no simple pattern for either N or L which could allow long recursion sequences to be predicted with confidence. Subsequently I looked at L for some much larger integers, taking sets of 200 consecutive N over the range from $N = 1000$ up to 160,200. The largest L in each group roughly approximated to $2 \cdot 1\sqrt{N}$.

3.5 Finding the square root of a fraction

§3.3 gave a scheme for finding the continued fraction of \sqrt{N} for N any positive integer. The method can be adapted to $\sqrt{N/M}$ for M and N any two positive integers. I'll give one example to illustrate the method and its differences from the single integer case. Of course an alternative method is always to determine $\sqrt{N/M}$ numerically, say with a hand calculator or by Newton's method, and then find the continued fraction for the resulting rational approximation by Euclid's method, §2.

Example : Determine $\sqrt{17/5}$ as a continued fraction. Since $1 < \sqrt{3 \cdot 4} < 2$, $a_0 = 1$ so write

N	L	N	L	N	L	N	L
7	4	76	12	134	14	214	26
13	5	86	10	139	18	244	26
19	6	93	10	151	20	508	32
21	6	94	16	157	17	811	38
22	6	97	11	163	18	1201	51
31	8	103	12	166	22	1471	64
43	10	109	15	172	16	1789	67
46	12	124	16	191	16	1831	84
61	11	127	12	199	20	1999	84
67	10	133	16	211	26	2011	94

Table 7: Lengths L of the longer recurring sequences for \sqrt{N} .

$\sqrt{17/5} = 1 + \delta$. Multiplying out

$$\text{Step 1:} \quad 5(1 + \delta)^2 = 17 \quad \text{from which} \quad 5\delta(2 + \delta) = 10\delta + 5\delta^2 = 12. \quad (3.7a)$$

Step 2: Now extract the integer part of $1/\delta$. From Eq 3.7a

$$\frac{1}{\delta} = \frac{10 + 5\delta}{12}.$$

The first difference from the simple \sqrt{N} case is that we need a fair estimate of the value of δ so that we can decide whether 5δ has an integer component. (In §3.3 multiples of δ do not appeared.) $18^2 = 324$ and $19^2 = 361$ so δ must be about 0.85 , making 5δ about 4.25 . The integer part of $(10 + 5\delta)/12$ is therefore $1 (= a_1)$ and we have

$$\frac{1}{\delta} = 1 + \frac{(-2 + 5\delta)}{12}. \quad (3.7b)$$

This presents the second difference from the simple \sqrt{N} case; the consequence of extracting an integer from 5δ is that negative values must be admitted into the calculation (but not into the a_k which are always positive in a simple continued fraction.)

Step 3: Now extract the integer part of $12/(-2 + 5\delta)$. The ‘trick’ in §3.3 was to replace a fraction having δ in the denominator by one having δ in the numerator. That is what Tables 4 and 5 are about. We must do the same here, so find constants A and B so that

$$\frac{12}{-2 + 5\delta} = \frac{A + \delta}{B} \quad \text{making} \quad 12B = -2A + (-2 + 5A)\delta + 5\delta^2.$$

The next difference from the single N case is how to make the δ^2 term disappear. From Eq 3.7a we know that $10\delta + 5\delta^2 = 12$, so choose A so that $-2 + 5A = 10$. Thus $A = 12/5$ and $B = 3/5$ and

$$\frac{12}{-2 + 5\delta} = \frac{5}{3} \left(\frac{12}{5} + \delta \right) = 4 + \frac{5\delta}{3} = 5 + \frac{(-3 + 5\delta)}{3}, \quad \text{and } a_2 = 5. \quad (3.7c)$$

Step 4: Continue in the same way, replacing $(-3 + 5A')\delta + 5\delta^2$ by $10\delta + 5\delta^2 = 12$:

$$\frac{3}{-3 + 5\delta} = \frac{A' + \delta}{B'} = \frac{5}{7} \left(\frac{13}{5} + \delta \right) = 2 + \frac{(-1 + 5\delta)}{7}, \quad \text{and } a_3 = 2. \quad (3.7d)$$

Step 5: Make $(-1 + 5\delta) = 10$:

$$\frac{7}{-1 + 5\delta} = \frac{5}{7} \left(\frac{11}{5} + \delta \right) = 2 + \frac{(-3 + 5\delta)}{7}, \quad \text{and } a_4 = 2. \quad (3.7e)$$

The fractions here are very similar to Eq 3.7c, and this is because the chain of partial quotients has started to reverse and retrace its steps. This is confirmed at the next three steps, which can be written down straight away by inspection of Eq 3.7 c), b) and a).

Step 6:

$$\frac{7}{-3 + 5\delta} = \frac{13 + 5\delta}{3} = 5 + \frac{(-2 + 5\delta)}{3}, \quad \text{and } a_5 = 5. \quad (3.7f)$$

Step 7:

$$\frac{3}{-2 + 5\delta} = 1 + \frac{5\delta}{12}, \quad \text{and } a_6 = 1. \quad (3.7g)$$

Step 8:

$$\frac{12}{5\delta} = 2 + \delta, \quad \text{so } a_7 = 2 \quad (3.7h)$$

and, with an isolated δ , we are back to the beginning of the recursion sequence. We have determined that $\sqrt{17/5} = \{1: \underline{1, 5, 2, 2, 5, 1, 2}\}$.

3.6 Pure square roots

We know that any recurring continued fraction must evaluate to a quadratic surd, but the question arises whether every continued fraction with the symmetric form of Eq 3.6 must evaluate to a pure square root, with no added integer. The answer is ‘yes’ though in general it is the square root of a rational number rather than of an integer, as we saw in the previous section, §3.5. I will give a proof for the case $\{ a : b, c, d, e, f \}$ since this illustrates the general case. Let the recurring part of this be θ as previously, and evaluate as a common fraction:

$$\begin{aligned} \frac{1}{e + \frac{1}{f+\theta}} &= \frac{f + \theta}{e(f + \theta) + 1} \\ \frac{1}{d + \frac{1}{e + \frac{1}{f+\theta}}} &= \frac{e(f + \theta) + 1}{(de + 1)(f + \theta) + d} \\ \frac{1}{c + \frac{1}{d + \frac{1}{e + \frac{1}{f+\theta}}}} &= \frac{(de + 1)(f + \theta) + d}{(cde + c + e)(f + \theta) + cd + 1} \\ \theta = \frac{1}{b + \frac{1}{c + \frac{1}{d + \frac{1}{e + \frac{1}{f+\theta}}}}} &= \frac{\theta(cde + c + e) + cde f + cd + cf + ef + 1}{\theta(bcde + bc + be + de + 1) + bcde f + bcd + bcf + bef + b + def + d + f} \end{aligned}$$

Multiplying out the denominator gives the quadratic equation

$$A\theta^2 + B\theta + C = 0$$

where

$$\begin{aligned} A &= bcde + bc + be + de + 1 \\ B &= bcdef + bcd + bcf + bef + b - cde - c + def + d - e + f \\ C &= -(cdef + cd + cf + ef + 1) \end{aligned}$$

There are two points to note. First, the palindromic symmetry : if $e = b$ and $d = c$, the single-factor terms in coefficient B cancel and make B exactly f times A :

$$A \rightarrow b^2(c^2 + 1) + 2bc + 1 \quad \text{and} \quad B \rightarrow f[b^2(c^2 + 1) + 2bc + 1].$$

Second, the solution of the quadratic for θ is

$$\frac{-B}{2A} \pm \frac{\sqrt{B^2 - 4AC}}{2A}.$$

Setting $f = 2a$ ensures that the integer part cancels, leaving only the square root. This proves that $\{ a : \underline{b, c, c, b, 2a} \}$ is always a pure square root of either an integer or a rational number.

Incidentally, C is then $-(2ab + 1)c^2 - 2a(b + c) - 1$.

3.7 Conjugate-simple continued fractions

In §1 we defined a simple continued fraction as one in which all the numerators are 1 and all the signs are +. Several other types of continued fraction can be formed by relaxing these conditions, and this is a suitable point to introduce ‘conjugate simple’ continued fractions. This is my own name for ones with 1 as each numerator, and all denominators of the form $a_k - 1/\dots$. I am calling them conjugate by analogy with algebraic and complex conjugates in which each + sign is swapped to a - sign. The notation $\{b_0 : b_1, b_2, b_3, \dots\}^*$ means

$$b_0 - \frac{1}{b_1 - \frac{1}{b_2 - \frac{1}{b_3 - \dots}}}.$$

Their convergents follow the recurrence relations

$$\frac{p_k}{q_k} = \frac{b_k p_{k-1} - p_{k-2}}{b_k q_{k-1} - q_{k-2}}, \quad (3.8)$$

similar to Eq 1.2 but with a + sign instead of -. They differ from the corresponding simple continued fraction by forming a strictly decreasing sequence, approaching their limiting value C_F monotonically. (Recall that simple continued fractions are alternating.)

In the simplest case where we have only one recurring partial quotient, b , there is an interesting relation between the conjugate-simple fraction and another type of non-simple fraction in which the signs are + but the partial quotients are square roots. To see this let

$$\psi = \{b : \underline{b}\}^* = b - \frac{1}{b - \frac{1}{b - \dots}}.$$

Then

$$\psi = b - \frac{1}{\psi} \quad \text{or} \quad \psi^2 - b\psi + 1 = 0$$

with solutions

$$\psi = \frac{1}{2}(b + \sqrt{b^2 - 4}) \quad \text{and} \quad \frac{1}{\psi} = \frac{1}{2}(b - \sqrt{b^2 - 4}).$$

Now compare this with a simple continued fraction

$$\theta = \{a : \underline{a}\} = a + \frac{1}{a + \frac{1}{a + \dots}} .$$

Then

$$\theta = a + \frac{1}{\theta} \quad \text{or} \quad \theta^2 - a\theta - 1 = 0$$

with solutions

$$\theta = \frac{1}{2}(a + \sqrt{a^2 + 4}) \quad \text{and} \quad \frac{1}{\theta} = \frac{1}{2}(a - \sqrt{a^2 + 4}).$$

θ will equal ψ if $a = \sqrt{b^2 - 4}$ and $b = \sqrt{a^2 + 4}$. We arrive at the relations

$$\{\underline{b}\}^* = \{\sqrt{b^2 - 4}\} \quad \text{and} \quad \{\underline{a}\} = \{\sqrt{a^2 + 4}\}^* . \quad (3.9)$$

As an example, the Golden Mean as a simple continued fraction is $\{\underline{1}\}$, with alternating convergents

$$1, \quad 2, \quad \frac{3}{2}, \quad \frac{5}{3}, \quad \frac{8}{5}, \quad \frac{13}{8}, \quad \frac{21}{13} = 1.6154, \quad \dots ,$$

while the conjugate-simple equivalent is $\{\sqrt{5}\}^*$ with monotonically decreasing convergents

$$\sqrt{5}, \quad \frac{4}{5}\sqrt{5}, \quad \frac{3}{4}\sqrt{5}, \quad \frac{11}{15}\sqrt{5}, \quad \frac{8}{11}\sqrt{5}, \quad \frac{29}{40}\sqrt{5} = 1.6211, \quad \dots .$$

3.8 Generalised continued fractions

Let me introduce *generalised* continued fractions, characterised by the numerator not being restricted to 1. The structure is

$$n_0 + \frac{n_1}{d_1 + \frac{n_2}{d_2 + \frac{n_3}{d_3 + \frac{n_4}{\dots}}}} \quad (3.10)$$

where n denotes numerator and d denominator. Several authors would denote this as

$$\left\{ n_0 + \frac{n_1}{d_1 +} \frac{n_2}{d_2 +} \frac{n_3}{d_3 +} \dots \right\}$$

However, I propose to use the double brace notation

$$\left\{ \left\{ n_0 : \frac{n_1}{d_1}, \frac{n_2}{d_2}, \frac{n_3}{d_3}, \dots \right\} \right\} .$$

This will mean that a simple continued fractions which we have previously notated as $\{a : b, c, d, \dots\}$ could be written as $\{\{a : 1/b, 1/c, 1/d, \dots\}\}$.

The generalised continued fraction Eq (3.9) obeys a recursion relation which is a generalisation of Eq 1.2 in §1.4:

$$p_{k+1} = d_{k+1}p_k + n_{k+1}p_{k-1} \quad \text{and} \quad q_{k+1} = d_{k+1}q_k + n_{k+1}q_{k-1}, \quad (3.11)$$

where $d_0 = 1$. In the matrix notation of §1.4.1 this is

$$\begin{pmatrix} d_k & n_k \\ 1 & 0 \end{pmatrix} \begin{pmatrix} p_{k-1} & q_{k-1} \\ p_{k-2} & q_{k-2} \end{pmatrix} = \begin{pmatrix} d_k p_{k-1} + n_k p_{k-2} & d_k q_{k-1} + n_k q_{k-2} \\ p_{k-1} & q_{k-1} \end{pmatrix} = \begin{pmatrix} p_k & q_k \\ p_{k-1} & q_{k-1} \end{pmatrix} .$$

For example

$$\begin{pmatrix} d_2 & n_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} d_1 & n_1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} n_0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} d_2(d_1n_0 + n_1) + n_2n_0 & d_2d_1 + n_2 \\ d_1n_0 + n_1 & d_1 \end{pmatrix} = \begin{pmatrix} p_2 & q_2 \\ p_1 & q_1 \end{pmatrix}.$$

The determinant of this matrix is $-n_2n_1$, and similarly the generalisation of Eq 1.5 is

$$p_kq_{k-1} - p_{k-1}q_k = (-1)^{k-1}n_kn_{k-1} \dots n_2n_1. \quad (3.12)$$

From this the difference of adjacent convergents is (compare with Eq 1.6)

$$\Delta_k = C_k - C_{k-1} = \frac{(-1)^{k-1}n_kn_{k-1} \dots n_2n_1}{q_kq_{k-1}}. \quad (3.13)$$

The main point at this stage is that all recurring generalised continuous fractions, like their simple counterparts, evaluate to quadratic surds. Moreover, the extra freedom in the numerators means that some square roots look much neater when expressed as generalised fractions. For example, as a simple continued fraction, $\sqrt{31}$ is $\{5 : \underline{1, 1, 3, 5, 3, 1, 1, 10}\}$, but with a numerator of 6 it can be written as $\{6 : \underline{6/10}\}$. $\sqrt{46}$ is even longer, at $\{6 : \underline{1, 3, 1, 1, 2, 6, 2, 1, 1, 3, 1, 12}\}$, but shrinks to $\{6 : \underline{10/12}\}$. This is because

$$\theta = \frac{n}{d+\theta} \text{ has solution } \theta = \frac{1}{2}(\sqrt{d^2 + 4n} - d).$$

This formula can be applied perversely to produce other recurring continued fractions for the same quadratic, such as $\sqrt{31} = \{4 : \underline{15/8}\} = \{6 : \underline{-5/12}\}$. This does demonstrate that these representations are not unique, whereas simple continued fractions are always unique. One notable generalised continued fraction is $\{a : \underline{b/(2a)}\} = \sqrt{a^2 + b}$. Use is made of this in §6.4 to solve a Pell-like diophantine equation. We will also return to generalised continued fractions in Part V.

4 Non-quadratic irrationals and Lagrange's method

4.1 Cube and higher roots

Forming a continued fraction depends critically on being able to determine the integer part of any given improper fraction. If we are using a continued fraction method to determine the value of a root, we cannot *a priori* assume the value of that root in carrying out our working. With square roots we made use of improper fractions of the form $(K + \delta)/m$ for integers K and m and δ strictly < 1 . On dividing K by m the remainder is at most $m - 1$ so $(m - 1 + \delta)/m$ is strictly < 1 . The integer part could be determined by inspection without ambiguity.

Unfortunately, the equivalent expansion for cube and higher roots does not allow such ready, confident determination. Consider, for example, trying to determine the continued fraction for $\sqrt[3]{N}$ without knowing its precise numerical value. Writing $N = (n + \delta)^3$, the binomial expansion is $n^3 + 3n^2\delta + 3n\delta^2 + \delta^3$. So

$$N - n^3 = 3n(n + \delta)\delta + \delta^3, \text{ giving } \frac{1}{\delta} = \frac{3n(n + \delta) + \delta^2}{N - n^3}.$$

So what is the integer part of this? It depends on δ , whose value we do not yet know. The example of finding $\sqrt[3]{5}$ will make this clear. Here $n = 1$ so we start with

$$\frac{1}{\delta} = \frac{3(1 + \delta) + \delta^2}{4}.$$

Roughly, if $\delta < 1/3$, the integer part is zero, but if $\delta > 1/3$ the integer part is 1.

Still trying to evaluate $\sqrt[3]{5}$ purely by continued fractions, we might alternatively try the factorisation

$$u^3 - v^3 = (u - v)(u^2 + uv + v^2).$$

Clearly $a_0 = 1$. In calculating a_1 let $u = \phi = \sqrt[3]{5}$ to find

$$\frac{1}{\phi - 1} = \frac{1}{4}(\phi^2 + \phi + 1).$$

By evaluating the right side at the approximations $\phi = 1$ and $\phi = 2$ we find that the value must lie between $3/4$ and $7/4$. This is not sufficient to decide whether the integer part $a_1 =$ is 0, 1 or 2, so we need a better approximation. Using $5 = 135/27 \approx 125/27 = (5/3)^3$, we find that the continued fraction starts $\{1 : 1, \dots\}$. The next step requires the integer part of $4/(\phi^2 + \phi - 3)$. Since all relevant surds ϕ must lie in the ring $\mathbb{Q}[\sqrt[3]{5}]$ and be so a linear combination of $\phi^0 = \phi^3$, ϕ^1 and ϕ^2 , we can write

$$4 = (A\phi^2 + B\phi + C)(\phi^2 + \phi - 3)$$

and solve simultaneously for A , B and C . The result is

$$\frac{4}{(\phi^2 + \phi - 3)} = \frac{1}{3}(\phi^2 + 2\phi + 1).$$

At $\phi = 2$ the right side evaluates to 3 and at $\phi = 5/3$ evaluates to $64/27$, so $a_2 = 2$. Using this approach, $a_3 = 2$ also, so the continued fraction starts $\{1 : 1, 2, 2, \dots\}$. Though it is tedious to solve three simultaneous equations at every partial quotient, one could continue, using the last two convergents at each stage to decide the integer part – that is, the next a_k . The method would fail if these last two convergents gave answers slightly either side of some integer M since then one could

not be sure whether the next a_k is M or $M - 1$. In that event one might turn instead to numerical methods like Newton's.

It may have been these difficulties which prompted Lagrange to develop an iterative method for using continued fractions to solve polynomial equations by modifying the equation itself for each successive convergent. Here is an example of his method applied to $\sqrt[3]{5}$. Let z (real) satisfy $f_0(z) = z^3 - 5 = 0$. By testing with some integer values we easily find that this has a root between $z = 1$ and $z = 2$. So write $z = 1 + 1/y$ and substitute into the original equation. y must satisfy

$$\frac{-4y^3 + 3y^2 + 3y + 1}{y^3} = 0.$$

Call the numerator of this $f_1(y)$. Then $f_1(y) = 0$. We find that $f_1(1) = 3$ and $f_1(2) = -13$, so the integer part of y is also 1. Hence at the next iteration we substitute $y = 1 + 1/x$, and then have $f_2(x) = 3x^3 - 3x^2 - 9x - 4 = 0$. This has a root between $x = 2$ and 3. By this stage our continued fraction approximation to $\sqrt[3]{5}$ is

$$1 + \frac{1}{1 + \frac{1}{2}} = \frac{5}{3}$$

to compare with the true value of $1.709976\dots$. The process can be continued through the chain of equations, substitutions and integer parts of roots listed in Table 8. Notice the coefficient's alternating signs. We have obtained $\sqrt[3]{5}$ as its convergent $\{1 : 1, 2, 2, 4, 3, 3\} = 566/331 = 1.709970$.

Substitution	Equation	Integer part of root
	$z^3 - 5 = 0$	1
$z = 1 + 1/y$	$-4y^3 + 3y^2 + 3y + 1 = 0$	1
$y = 1 + 1/x$	$3x^3 - 3x^2 - 9x - 4 = 0$	2
$x = 2 + 1/w$	$-10w^3 + 15w^2 + 15w + 3 = 0$	2
$w = 2 + 1/v$	$13v^3 - 45v^2 - 45v - 10 = 0$	4
$v = 4 + 1/u$	$-78u^3 + 219u^2 + 111u + 13 = 0$	3
$u = 3 + 1/t$	$211t^3 - 681t^2 - 483t - 78 = 0$	3

Table 8: Lagrange's method for $\sqrt[3]{5}$

How does this compare with, say, Newton's method for finding the roots of an equation? Its main advantage is that the intermediate equations need to be evaluated only at integers. It does give the simple continued fraction directly, and we know that each convergent is the best rational approximation to the root of any fraction with equal or smaller denominator. On the disadvantage side, it requires increasing algebraic manipulation to obtain a new equation for every convergent. Moreover, there must be doubt that it would find all roots of a polynomial, particularly if two lie close together, between the same two integers. In view of this it may be easier merely to evaluate $\sqrt[n]{N}$ numerically by a standard method such as Newton's to the desired number of decimal places, then treat it as a rational number and find its continued fraction by the methods of §2. In the case of $\sqrt[3]{5}$ this is $1709976/1000000$ etc. using high precision computational software. The result is $\{1 : 1, 2, 2, 4, 3, 3, 1, 5, 1, 1, 4, 10, 17, 1, 14, 1, 1, 3052, 1, \dots\}$ which, as expected, shows no evidence of

recurring. Notice also the large partial quotient 3052. The probability of any given value v amongst the a_k is explored in Part IV.

4.2 Calculations for transcendental numbers

In most cases calculating the continued fraction expansion of a transcendental number (*i.e.* one which does not satisfy a polynomial equation) is possible only if you have a sufficiently accurate decimal approximation. You then just treat it as a common fraction and evaluate by the methods of §2. There are few special methods for transcendentals. However, one type which can be calculated in continued fractions is the logarithm of a given real. The algorithm for this is described in §5, Part II, as an example of a practical application of continued fractions. The method was once well known but now is probably forgotten.

Take π as an example. Finding its continued fraction first requires a decimal approximation. One approach might be by repeated application of the trigonometric double angle formula

$$\cos 2\theta = 2 \cos^2 \theta - 1, \text{ from which } \cos \frac{\theta}{2} = \sqrt{\frac{1 + \cos \theta}{2}}.$$

Starting with $\cos \frac{\pi}{2} = 0$ and taking repeated roots, $\cos(\pi/1024) = 0.99999529381$. Now $\cos \theta \approx 1 - \theta^2/2$. Therefore $\pi \approx 3 \cdot 141591$, with corresponding continued fraction $\{ 3 : 7, 15, 1, 39, \dots \}$. Another approach would be Machin's formula

$$\frac{\pi}{4} = \arctan\left(\frac{1}{5}\right) - \arctan\left(\frac{1}{239}\right)$$

or one of the more powerful modern Machin-like formulae. With a more precise evaluation of π the continued fraction begins $\{ 3 : 7, 15, 1, 292, 1, 1, \dots \}$. The first convergent is $22/7$, the traditional approximation once taught in primary schools. The higher convergent $355/113$ gives π correct to better than 3 parts in 10^7 .

An alternative approach to π might be to use Lagrange's equation-changing method to solve an equation for π . So let's apply this to $\sin(\pi/4) = 1/\sqrt{2}$. In principle all we have to do is determine the two integers between which the current expression changes sign. Table 9 gives the sequence of equations, the integer part of the root, and the implied substitution which lead to the next equation. We thereby arrive at $\frac{\pi}{4} = \{0 : 1, 3, 1, 1, 1, 15, 2, 72, 1, 9, 1, \dots\}$. However, to determine that the integer part of s is 2 in the above table requires that we have sine tables accurate enough to decide that there is a root of $\sin(z) = 1/\sqrt{2}$ between $z = 0.7853982$ ($s = 2$) and $z = 0.7853949$ ($s = 3$). Whether values of $\sin z$ would be known any better than the value of π is a moot point, so the practical value of Lagrange's method is also open to debate. In Lagrange's day, however, it seems to have been regarded as an important addition to the tool kit of numerical computation.

The transcendental number $e \approx 2.71828$ is particularly interesting because its continued fraction has a strong pattern: $\{ 2 : 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, 1, 1, 12, 1, 1, 14, 1, 1, 16, \dots \}$. The pattern is even more marked for the constant $1/(e-1) = \{0 : 1, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, \dots\}$. This was found by Euler. It was a step towards a proof that e is transcendental (though transcendental numbers were not known in Euler's time). They argued correctly that because the continued fraction is infinite, e cannot be rational, and because it is non-recurring, it cannot be a square root. Moreover, like the fraction for π , the a_k have no upper bound – this is a key to transcendental behaviour, as will be explored in Part III, §12.

Table 9: Lagrange's method for $\sin^2 z = \frac{1}{2}$.

Substitution	Equation	Integer part of root
	$\sin^2 z = \frac{1}{2}$	0
$z = 1/y$	$\sin^2 \frac{1}{y} = \frac{1}{2}$	1
$y = 1 + 1/x$	$\sin^2 \frac{x}{x+1} = \frac{1}{2}$	3
$x = 3 + 1/w$	$\sin^2 \frac{3w+1}{4w+1} = \frac{1}{2}$	1
$w = 1 + 1/v$	$\sin^2 \frac{4v+3}{5v+4} = \frac{1}{2}$	1
$v = 1 + 1/u$	$\sin^2 \frac{7u+1}{9u+5} = \frac{1}{2}$	1
$u = 1 + 1/t$	$\sin^2 \frac{11t+1}{14t+9} = \frac{1}{2}$	15
$t = 15 + 1/s$	$\sin^2 \frac{172s+11}{219s+14} = \frac{1}{2}$	2

It is clearly easy to invent continued fractions whose partial quotients follow some pattern. In a few cases these can be identified with tabulated constants. Moreover, by making the partial quotient a variable, x say, one can invent functions of x (provided the continued fraction converges). There have been some remarkable results found by Euler, Gauss and others for functions expressed as continued fractions, and some of these are outlined in Part IV.

Before leaving this subsection, let me just quote a remarkable series for π discovered by Ramanujan in about 1910 (or divinely revealed to him, as he would have explained). This is an example of approximating a transcendental by an algebraic irrational. Since π is transcendental, it has an infinite tail to its representation either as a decimal or as a continued fraction. Rational approximations to π (which have a history back to Archimedes) can never represent the tail of the decimal, only a finite leading part. Ramanujan's formula uses another irrational, $\sqrt{2}$, as the kernel of the approximation, making use of the built-in infinite tail of this algebraic number. His formula is

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{9801} \sum_{n=0}^{\infty} \frac{(4n)!(1103 + 26390n)}{(n!)^4 396^{4n}}.$$

Convergence is astoundingly fast; the first term alone gives

$$\pi \approx \frac{9801}{2206\sqrt{2}}$$

which is in error by only $7 \cdot 6 \times 10^{-8}$. The fraction $9801/2206$ is actually convergent C_7 of $\pi\sqrt{2} = \{ 4 : 2, 3, 1, 7, 7, 1, 3, 1, \dots \}$. Of course, the great power of Ramanujan's formula is that it can be used to calculate π , but if you know π , it is easy to find other approximations similar to $9801/(2206\sqrt{2})$. Here are a few:

$$\begin{array}{ll} \pi \approx \frac{739}{700}\sqrt{11} & |\epsilon| = 5 \cdot 2 \times 10^{-7} \\ \frac{7442}{7491}\sqrt{10} & |\epsilon| = 3 \times 10^{-8} \\ \frac{9941}{8372}\sqrt{7} & |\epsilon| = 1 \times 10^{-8} \\ \frac{32864}{24695}\sqrt{11} & |\epsilon| = 1 \cdot 7 \times 10^{-10} \end{array}$$

Similarly you can readily find approximations to e involving an algebraic number, such as $699/(115\sqrt{5})$, which has error $9 \cdot 3 \times 10^{-7}$.

Ramanujan knew a remarkable rational approximation to π^4 . By a fluke of numbers the continued fraction for π^4 happens to be $\{97 : 2, 2, 3, 1, 16539, 6, 7, \dots\}$. The partial quotient 16539 makes such a small difference that the previous convergent $2143/22$ is accurate to 1 part in 10^8 . Ramanujan had a reverence for such numbers.

Part II

Some applications of continued fractions

We now have enough background on continued fractions to study some interesting applications. This Part deals with three: Bourdon's method for calculating logarithms (§5), Pell's equation (§6), and the CFRAC factorisation algorithm (§7). Another useful application is Thiele's algorithm for fitting a rational function to given data points, but this is best described in the context of continued fractions of functions $f(x)$, so is deferred until §18, Part IV.

5 Calculating logarithms

The method described here for determining the logarithm to a given base of a given positive real number was published in an algebra textbook, in French, by Marie Pierre Bourdon in 1828. The method, which is similar to Lagrange's method of §4.1, was probably well known at the time, but in our electronic, post-log-table times it has been almost forgotten.

Recall that the logarithm λ of a real number R to a base b , written $\log_b R = \lambda$, satisfies the exponential equation

$$b^\lambda = R. \quad (5.1)$$

If R is an integer power of b , the logarithm is just that integer, and there is nothing more to calculate. So suppose that $b^{a_0} < R$ but $b^{a_0+1} > R$. a_0 is an integer, easily calculated. By analogy with Euclid's method, and the basic algorithm for continued fractions, Eq 1.1, we extract the integer part of the power and then take the reciprocal of the remainder. So

$$b^{a_0} \cdot b^{\lambda-a_0} = b^{a_0} \cdot \frac{R}{b^{a_0}}, \quad \text{implying} \quad b^{\lambda-a_0} = \frac{R}{b^{a_0}} = R_1, \quad \text{say.}$$

Now $\lambda - a_0 < 1$, so its reciprocal is > 1 . Call this reciprocal λ_1 . Then

$$R_1^{\lambda_1} = b$$

which has the same form as Eq 5.1. The process can now be repeated, and repeated again, until the precision of the arithmetic is exhausted. Stage by stage the logarithm is developed as a continued fraction

$$\lambda = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \dots}}$$

An example should make this clear.

Example : Calculate $\log_{10} 200$.

Stage 1 : Start by noting that $10^2 = 100 < 200$ and $10^3 = 1000 > 200$ so $a_0 = 2$. Then

$$10^2 \cdot 10^{\lambda-2} = 10^2 \cdot 2 \quad \text{so} \quad 10^{\frac{1}{\lambda_1}} = 2, \quad 10 = 2^{\lambda_1},$$

where $\lambda_1 = 1/(\lambda - 2)$.

Stage 2 : $2^3 = 8 < 10$ whilst $2^4 = 16 > 10$ so $a_1 = 3$. Then

$$2^3 \cdot 2^{\lambda_1-3} = 2^3 \cdot \frac{10}{8} \quad \text{so} \quad 2^{\frac{1}{\lambda_2}} = \frac{5}{4}, \quad 2 = \left(\frac{5}{4}\right)^{\lambda_2},$$

where $\lambda_2 = 1/(\lambda_1 - 3)$.

Stage 3 : The fractions now become rather cumbersome. $(5/4)^3 = 125/64 < 2$, just, so $a_3 = 3$. Then

$$\left(\frac{5}{4}\right)^3 \left(\frac{5}{4}\right)^{\lambda_2-3} = \left(\frac{5}{4}\right)^3 \left(\frac{128}{125}\right) \quad \text{so} \quad \frac{5}{4} = \left(\frac{128}{125}\right)^{\lambda_3},$$

where $\lambda_3 = 1/(\lambda_2 - 3)$.

So far we have determined that $\lambda = \{2 : 3, 3, \dots\} = 23/10 = 2 \cdot 30$. In fact $10^{2 \cdot 3} = 199 \cdot 53$. Clearly $\log_{10} 2 = 0.30\dots$

This method must have been tedious for clerks with pen and ink to carry out by hand in the early nineteenth century, but it is straightforward in that it requires only multiplication and division. Today, on a computer furnished with arbitrary precision arithmetic software, one can obtain an indefinite number of convergents to the logarithm. For instance, I find that

$$\log_{10} 2 = \{0 : 3, 3, 9, 2, 2, 4, 6, 2, 1, 1, 3, 1, 18, 1, 6, 1, 2, 2, 21, 1, 3, \dots\}$$

with convergents

$$0, \quad \frac{1}{3}, \quad \frac{3}{10}, \quad \frac{28}{93}, \quad \frac{59}{196}, \quad \frac{146}{485}, \quad \text{etc.}$$

and value $0.3010299956\dots$. The error in the convergent $146/485$ above is only 1×10^{-6} .

Clearly the method can be used to solve numerically any equation of the form $a = b^c$.

6 Pell's equation

6.1 Pell's equation and its solution by continued fractions

We saw in §1.5 and Eq 1.5 that any pair of adjacent convergents can be cross multiplied and their products differ by ± 1 . In §2.2 this was used to solve linear Diophantine equations of the form $Mx \pm Ny = \kappa$, using x/y as a convergent of N/M . Now we ask 'Can we use continued fractions to solve equations in integers of the quadratic form

$$\text{Generalised Pell equation: } \quad Mx^2 - Ny^2 = \kappa \quad ? \quad , \quad (6.1a)$$

This equation is named after the 17th century Englishman John Pell, though he seems to have had little to do with it. It was investigated in detail by Lagrange and the great Euler. For Pell, this was another case of Reward of the Uninvolved! Pell's equation is often presented in its simplest form with $M = 1$ and $\kappa = \pm 1$: that is

$$\text{Basic Pell equation: } \quad x^2 - Ny^2 = \pm 1 \quad (6.1b)$$

where $N > 1$ is not a perfect square.

So, suppose you were asked to find integer solutions to $Mx^2 - Ny^2 = \kappa$ where κ is ± 1 or other small integer, but M and N could be large. (Without loss of generality, take $N > M$.) How to go about it? You might start with the simpler but similar equation $Mx^2 - Ny^2 = 0$, for which the solution would be $\frac{x}{y} = \sqrt{\frac{N}{M}}$. You would then be looking for pairs of integers which are approximately in the ratio of this square root. But of all such ratios x/y which might be tried, the convergents of $\sqrt{N/M}$ come closest, by their 'best fit' property. So the convergents of $\sqrt{N/M}$ are the natural first choice for solutions of Pell's equation.

Let's test this with a simple numerical example of the basic Pell equation. Take the first few convergents of $\sqrt{5} = \{ 2 : \underline{4} \}$

$$\sqrt{5} \approx \frac{2}{1}, \frac{9}{4}, \frac{38}{17}, \frac{161}{72}, \frac{682}{305}, \frac{2889}{1292}, \frac{12238}{5473}, \dots$$

Consistent with Eq 1.5, we see that $9 \times 17 - 4 \times 38 = 1$ and similarly for other adjacent pairs. But now note the additional behaviour of the numerator and denominator of each convergent:

$$\begin{aligned} 2^2 = 4 \quad \text{and} \quad 5 \times 1^2 = 5 \\ 9^2 = 81 \quad \text{and} \quad 5 \times 4^2 = 80 \\ 38^2 = 1444 \quad \text{and} \quad 5 \times 17^2 = 1445 \\ 161^2 = 25921 \quad \text{and} \quad 5 \times 72^2 = 25920 \\ 682^2 = 465124 \quad \text{and} \quad 5 \times 305^2 = 465125, \quad \text{etc.} \end{aligned}$$

with an alternating difference of -1 or $+1$. Without any deeper analysis we have found an infinite number of solutions in integers to the two basic Pell equations

$$x^2 - 5y^2 = 1 \quad \text{and} \quad x^2 - 5y^2 = -1$$

in which $M = 1$, $N = 5$ and κ is alternatively -1 , $+1$ as we progress along the sequence of convergents.

Clearly κ is related to the error in the k^{th} convergent $x/y = p_k/q_k$ to approximate $\sqrt{N/M}$. This is emphasised by writing Pell's equation in the form

$$\frac{p_k^2}{q_k^2} - \frac{N}{M} = C_k^2 - \frac{N}{M} = \frac{\kappa}{Mq_k^2}.$$

$$\text{Now } C_k^2 - \frac{N}{M} = \left(C_k - \sqrt{\frac{N}{M}}\right) \left(C_k + \sqrt{\frac{N}{M}}\right) \approx 2\epsilon_k \sqrt{\frac{N}{M}}. \quad (6.2)$$

In §1.7, Eq 1.13 we obtained the error bounds

$$\frac{1}{(a_{k+1} + 2)q_k^2} < \epsilon_k < \frac{1}{a_{k+1}q_k^2}$$

and later, in §11.4, we will obtain the estimate

$$\epsilon_k \approx \frac{1}{\sqrt{2}a_{k+1}q_k^2}.$$

Combining these

$$\frac{\kappa}{M} \approx \frac{1}{a_{k+1}} \sqrt{\frac{2N}{M}}. \quad (6.3)$$

which should apply in the majority of cases ⁴.

In the generalised Pell equation the convergents are not the only possible choices for x/y . In §2.3 and §9.3 I describe 'quasi-convergents' which are derived from authentic convergents and do give a modest improvement in accuracy over fractions with similar denominator. Take the above example of $x^2 - 5y^2 = \kappa$ in which $M = 1$. Three quasi-convergents of $\sqrt{5}$ are $29/13 = \{2: 4, 3\}$, $47/21 = \{2: 4, 5\}$ and $123/55 = \{2: 4, 4, 3\}$ all give $|\kappa| = 4$, while $85/38 = \{2: 4, 4, 2\}$ gives $\kappa = 5$. In each case κ is small, but not 1. However, in the particular case $|\kappa| = M$ x/y must be a convergent of $\sqrt{N/M}$. To see why consider that in Eq 5.2 $N > M$ by definition and C_k , being an approximation to $\sqrt{N/M}$, must also be > 1 . Therefore $C_k + \sqrt{N/M} > 2$, making

$$C_k^2 - \frac{N}{M} = \frac{1}{q_k^2} > 2\epsilon_k, \quad \text{so } \epsilon_k < \frac{1}{2q_k^2}.$$

From Eq 1.11 and its proof in §8.2, this is sufficient for x/y to be a convergent of $\sqrt{N/M}$. In our example only the authentic convergents are close enough approximations to $\sqrt{5}$ to give $\kappa = \pm 1$.

6.2 Patterns in κ for $x^2 - Ny^2 = \kappa$

The case of $x^2 - 5y^2 = \pm 1$ was particularly simple because all the partial quotients of $\sqrt{5}$ are 1. Let's see what patterns occur with a square root whose recurring sequence is longer. It happens that the behaviour depends on whether the length of the sequence, L , is odd or even, so I will give an example of each, setting $M = 1$. Following §3.4.3 remember that any continued fraction of the symmetric form $\{a : \underline{b, c, d, e, \dots, e, d, c, b, 2a}\}$ represents the pure square root of a rational number.

For $L = 3$ my example is $N = 41$ since $\sqrt{41} = \{6 : 2, 2, 12\}$ is one of the few square roots of an integer with a recursion length $L = 3$. Table 10 shows the convergents and, in the last column, the Pell equation formed from numerator and denominator. Notice how $|\kappa| = 1$ is obtained only at every

Table 10: Forms of Pell's equation based on convergents to $\sqrt{41} = \{6 : \underline{2, 2, 12}\}$

k	a_k	p_k	q_k	$\kappa = p_k^2 - 41q_k^2$
0	6	6	1	-5
1	2	13	2	5
2	2	32	5	-1
3	12	397	62	5
4	2	826	129	-5
5	2	2049	320	1
6	12	25414	3969	-5
7	2	52877	8258	5
8	2	131168	20485	-1
9	12	1626893	254078	5
10	2	3384954	528641	-5
11	2	8396801	1311360	1
12	12	104146566	16264961	-5
13	2	216689933	33841282	5
14	2	537526432	83947525	-1
15	12	6667007117	1041211582	5

third convergent, for $k = 2, 5, 8, 11, 14, \dots, 2n - 1$. Moreover, these values of κ alternate in sign. At the other convergents κ is -5 or $+5$, these values being symmetrically placed about $|\kappa| = 1$.

As an example of an L even case, take $\sqrt{19} = \{4 : \underline{2, 1, 3, 1, 2, 8}\}$. The convergents are listed by numerator and denominator in Table 11, and the last column shows κ . For all convergents κ is small and exactly $+1$ when the recurring sequence is truncated to include the second '2' partial quotient but not the '8', as marked by the † in the table. Values of κ alternate in sign and are symmetrical in the table about the value '1'.

If you were to find κ_k for the Pell equation $x^2 - 103y^2 = \kappa$, you would find the sequence $\kappa_0 = -3$, then 13, -6 , 9, -11 , 2, -11 , 9, *etc.*. A glance at §3.3, Table 5 and the example of finding $\sqrt{103}$, will show that the above κ_k values are precisely the denominators in the chain of factorisations. Pell's equation is intimately related to the procedure in §3.3 for finding the square root.

To summarise findings, numerical investigation using other square roots supports these observations:

1. the values $\kappa = \pm 1$ can be obtained only if $M = 1$,
2. if the recursion length is L , only the convergents C_{nL-1} give $|\kappa| = M$, for n any positive integer,
3. the other values of κ are symmetrically placed around positions $k = nL - 1$ at which $|\kappa| = M$,
4. if L is odd, the lowest value of k to give $|\kappa| = M$ is $k = L - 1$ and the value there is $\kappa = -M$. Thereafter the sign of κ alternates,

⁴In the special case of a single recurring partial quotient, $\sqrt{N/M} = \{a/2 : \underline{a}\}$, the error is given exactly by $\epsilon_k = 1/(\sqrt{(a^2 + 4)q_k^2})$. This is proved in §9.1.2. Applying this to the above case of $\sqrt{5} = \{2 : \underline{4}\}$ gives $\kappa = 1$ exactly.

Table 11: Forms of Pell's equation based on convergents to $\sqrt{19}$

k	a_k	$p_k = x$	$q_k = y$	x^2	$19y^2$	$x^2 - 19y^2$
0	4	4	1	16	19	-3
1	2	9	2	81	76	5
2	1	13	3	169	171	-2
3	3	48	11	2304	2299	5
4	1	61	14	3721	3724	-3
† 5	2	170	39	28900	28899	1
8	4	1421	326	2019241	2019244	-3
7	2	3012	691	9072144	9072139	5
8	1	4433	1017	19651489	19651491	-2
9	3	16311	3742	266048721	266048716	5
10	1	20744	4759	430313536	430313539	-3
† 11	2	57799	13260	3340724401	3340724400	1
12	8	483136	110839	233420394496	233420394499	-3
13	2	1024071	234938	1048721413041	1048721413036	5
14	1	1507207	345777	2271672940849	2271672940851	-2

5. if k is even, only positive values of $\kappa = +M$ are obtained. The value $\kappa = -M$ does not occur and the corresponding Pell equation has no solution.

6.2.1 The special qualities of Pell's equation being quadratic

We should not be surprised that the difference is small *relative* to x and y since x/y becomes a better approximation to $\sqrt{N/M}$ as we move to higher convergents. We probably should not be surprised at the cyclic values of the difference, which must arise in some way from the recurring sequence of partial quotients and their palindromic symmetry. But I do find it remarkable that the cyclic pattern ensures that $x^2 - \frac{N}{M}y^2$ remains small in *absolute* value, becoming as small as 1, even when x and y have tens of digits! This comes about because the error in a convergent, ϵ_k , decreases as $1/q_k^2$, exactly to meet the requirement on error if κ/M is to be constant in Pell's quadratic equation.

To bring this point home, consider the analogous challenge of finding integer solutions of the equation $Mx^3 - Ny^3 = \pm\kappa$ for κ small. Take the case $M = 1$, for simplicity, and $N = 5$. We might go about this by using Newton's method to find $\sqrt[3]{5}$, then its continued fraction, $\{1 : 1, 2, 2, 4, 3, 3, 1, 5, 1, 1, 4, 10, 17, \dots\}$, and evaluate successive convergents. The resulting numbers would be those in Table 12.

It is clear that, in absolute value, the difference in the last column gradually increases, decreasing only occasionally and never to an integer close to 1. Analysis of error similar that above predicts that in $Mx^3 - Ny^3 = \kappa$, κ should be roughly proportional to q_k/a_{k+1} ⁵. The higher power, 3,

⁵Let $x = p_k = p$, $y = q_k = q$.

$$\frac{\kappa}{Mq^3} = \frac{p^3}{q^3} - \frac{N}{M} = \left(\frac{p}{q} - h\right)\left(\frac{p^2}{q^2} + \frac{ph}{q} + h^2\right) \approx 3\epsilon h^2.$$

where $h = \sqrt[3]{N/M}$. Hence $\kappa_k/M \approx 2h^2 q_k/a_{k+1}$. This fits the trend in Table 12 fairly well.

Table 12: Solution of $x^3 - 5y^3 = \pm\kappa$ by convergents of $\sqrt[3]{5}$

$p_k = x$	$q_k = y$	x^3	$5y^3$	$x^3 - 5y^3$
2	1	8	5	3
5	3	125	135	-10
12	7	1728	1715	13
53	31	148877	148955	-78
171	100	5000211	5000000	211
566	331	181321496	181323455	-1959
737	431	400315553	400314955	598
4251	2486	76819825251	76819836280	-11029
4988	2917	124102158272	124102146065	12207
9239	5403	788632918919	788632929135	-10216
41944	24529	73792042960384	73792042939445	20939

magnifies the discrepancies between the approximating fraction and $\sqrt[3]{5}$, so making greater demands for smallness of error than the convergents can provide. So we should not be surprised that such cubic and higher power Diophantine equations have few solutions for κ small. Nevertheless, I did stumble across these few exceptions:

$$8^3 - 19 \times 3^3 = -1, \quad 18^3 - 17 \times 7^3 = 1, \quad 10^3 - 37 \times 3^3 = 1, \quad \text{and} \quad 467^3 - 6 \times 257^3 = 5.$$

These sporadic examples emphasise how the square root is peculiar in periodically and indefinitely returning very small integer differences in Pell's equation.

6.3 Some analysis of the generalised Pell equation $Mx^2 - Ny^2 = \kappa$

I will first analyse a generalised Pell equation for an $L = 2$ recursion sequence, namely $\sqrt{N/M} = \{a: b, 2a\}$. The recursion formula defines τ , the fractional part of $\sqrt{N/M}$, according to

$$\tau = \frac{1}{b + \frac{1}{2a + \tau}} \quad \text{from which} \quad \tau = -a + \sqrt{a^2 + \frac{2a}{b}}, \quad \theta = \sqrt{a^2 + \frac{2a}{b}}.$$

Hence

$$M = \frac{b}{g}, \quad N = \frac{a^2b + 2a}{g} \quad \text{where} \quad g = \gcd(a^2b + 2a, b).$$

If $b \nmid a$, we can take $M = b$ and $N = a^2b + 2a$.

The first few convergents of $\{a: b, 2a\}$ are

$$C_0 = a, \quad C_1 = \frac{ab + 1}{b}, \quad C_2 = \frac{2a^2b + 3a}{2ab + 1}, \quad C_3 = \frac{2a^2b^2 + 4ab + 1}{2b(ab + 1)}.$$

We form the generalised Pell equation for each convergent. Here are the equations and the values of κ for each:

1. $C_0 : b \cdot a^2 - (a^2b + 2a) = -2a,$

2. $C_1 : b.(ab + 1)^2 - (a^2b + 2a).b^2 = b = M,$
3. $C_2 : b.(2a^2b + 3a)^2 - (a^2b + 2a).(2ab + 1)^2 = -2a,$
4. $C_3 : b.(2a^2b^2 + 4ab + 1)^2 - (a^2b + 2a).(2b(ab + 1))^2 = b.$

Clearly κ is either b or $-2a$, alternating in sign and between the two partial quotients in the recursion sequence. We will prove shortly that this pattern alternates indefinitely. Numerical values bear this out. For example, $a = 1, b = 5$ gives $M = 5, N = 7$. The first five convergents are $1, 6/5, 13/11, 71/60, 155/131$, and they give respectively κ to be $-2, 5, -2, 5, -2$.

The next recursion sequence in complexity is $\{a : \underline{b, b, 2a}\}$. Results equivalent to the case above are as follows.

$$M = b^2 + 1, \quad N = a^2(b^2 + 1) + 2ab + 1.$$

$$C_0 = a, \quad C_1 = \frac{ab + 1}{b}, \quad C_2 = \frac{ab^2 + a + b}{b^2 + 1}, \quad C_3 = \frac{2a^2(b^2 + 1) + 3ab + 1}{2a(b^2 + 1) + b},$$

$$C_4 = \frac{2a^2b^3 + 2a^2b + 4ab^2 + a + 2b}{2ab(b^2 + 1) + 2b^2 + 1}.$$

1. $C_0 : \kappa_0 = -(2ab + 1)$
2. $C_1 : \kappa_1 = 2ab + 1 = N - a^2M$
3. $C_2 : \kappa_2 = -(b^2 + 1) = -M$
4. $C_3 : \kappa_3 = 2ab + 1 = \kappa_1$
5. $C_4 : \kappa_4 = -(2ab + 1) = \kappa_0$, etc.

As a numerical example take $a = 1, b = 3$. Then $M = 10, N = 17$, and so Pell's equation is $10x^2 - 17y^2 = \kappa$. The respective values of κ_0 to κ_8 are $-7, 7, -10, 7, -7, 10, -7, 7, -10$. Signs alternate, modulating the repeated three-fold pattern $\{7, 7, 10\}$. Therefore the pattern in κ has a repeat length of 6. In this example $2a < b$ so the lowest value of $|\kappa|$ is $N - a^2M = 7$, not $M = 10$.

The complexity of the expressions for M, N and κ increase significantly for longer recursion sequences, so the last one I will present is $\{a : \underline{b, c, b, 2a}\}$. Results equivalent to the cases above are:

$$M = b(bc + 2), \quad N = (ab + 1)(abc + 2a + c)$$

1. $C_0 : \kappa_0 = -[2a(bc + 1) + c],$
2. $C_1 : \kappa_1 = 2b(ab + 1),$
3. $C_2 : \kappa_2 = \kappa_0,$
4. $C_3 : \kappa_3 = b(bc + 2) = M,$
5. After this $\kappa_4 = \kappa_0, \quad \kappa_5 = \kappa_1, \quad \kappa_6 = \kappa_0, \quad \kappa_7 = \kappa_3 = M,$

and the cycle repeats with $\kappa_j = \kappa_{j-4}$. An example of this $L = 4$ sequence is $a = 1, b = 3, c = 4$ which gives $M = 42, N = 72$ with $\gcd(M, N) = 6$, making $N/M = 12/7$. Divide the Pell equation through by 6 to find that $7p_k^2 - 12q_k^2 = \kappa'_k = \kappa_k/6$. For C_0 to C_6 , κ'_k is $-5, 4, -5, 7, -5, 4, -5$.

Notice the alternating sign, and how in all cases the value $\kappa = M$ is obtained from the penultimate partial quotient in the sequence, immediately before the $2a$. Here is a proof of this fact. Consider the pure square root $\theta = \sqrt{N/M} = \{a : \underline{b, c, d, \dots, d, c, b, 2a}\}$ and τ be the fractional part $\{\underline{b, c, d, \dots, d, c, b, 2a}\}$. This means that $\theta = a + \tau$. Also let the convergent $C_k = p_k/q_k = \{a : \underline{b, c, d, \dots, d, c, b, 2a, b, c, d, \dots, d, c, b}\}$; that is, it corresponds to truncating the repeated pattern of partial quotients immediately before any one of the $2a$. From the exact recursion relation Eq 1.3,

$$\theta = \frac{(2a + \tau)p_k + p_{k-1}}{(2a + \tau)q_k + q_{k-1}} = \frac{(a + \theta)p_k + p_{k-1}}{(a + \theta)q_k + q_{k-1}}. \quad (6.4)$$

This is a quadratic equation for θ :

$$q_k \theta^2 + (aq_k + q_{k-1} - p_k)\theta - (ap_k + p_{k-1}) = 0 \quad \text{or} \quad \mathcal{A}\theta^2 + \mathcal{B}\theta + \mathcal{C} = 0.$$

Since θ is a pure square root, the coefficient \mathcal{B} must be zero; that is

$$aq_k + q_{k-1} = p_k \quad \text{and} \quad \theta = \sqrt{\frac{-\mathcal{C}}{\mathcal{A}}}. \quad (6.5)$$

so we can identify M with $\mathcal{A} = q_k$ and N with $-\mathcal{C} = ap_k + p_{k-1}$. The final stage is to form the Pell equation from convergent C_k . To evaluate κ use Eq 6.4 and Eq 1.5:

$$\begin{aligned} Mp_k^2 - Nq_k^2 &= q_k p_k^2 - (ap_k + p_{k-1})q_k^2 \\ &= q_k [p_k(aq_k + q_{k-1}) - ap_k q_k - p_{k-1} q_k] \\ &= q_k (-1)^{k-1} \\ &= (-1)^{k-1} M. \end{aligned}$$

C_k was defined by truncating the continued fraction for θ immediately before *any* of the partial quotients $2a$ and there are an infinite number of these whenever $k = \lambda L - 1$, for any positive integer λ , where L is the length of the recursion sequence. If L is odd, the values of k and hence κ are alternately odd and even. However, if L is even, k is always odd and $\kappa - 1$ is always even, so there are no solutions $\kappa = -M$, consistent with point 5) in the list of §6.2.

Regarding values of κ when θ is truncated before partial quotients other than $2a$, the examples above show that κ will depend in a non-trivial way on M and N , that is, on the precise recursion sequence, and in some cases $|\kappa|$ can be less than M .

The expression $Mx^2 - Ny^2$ in Pell's equation is an example of a 'binary quadratic form', 'binary' meaning that there are two variables and 'quadratic' that it is of degree 2. The most general binary quadratic form is $AX^2 + BXY + CY^2$, denoted $\mathcal{F}(A, B, C)$. If $\mathcal{F}(A, B, C) = \kappa$ for some integer κ , as in Pell's equation, we say that κ can be represented by the form \mathcal{F} . Quadratic and higher forms were studied extensively in the late 18th century and 19th century by Gauss and others, in a search for general algorithms for solving Diophantine equations. There is a deep link between the properties of binary quadratic forms and continued fractions; some introductory aspects will be described in §12.4.

6.4 Another Pell-like diophantine equation

Integer solution sets (x, y, z, n) to the diophantine equation

$$x^2 - Ny^2 = z^n \quad (6.6)$$

can be found using the recurring generalised continued fraction introduced at the end of §3.8, namely

$$\{a : \underline{b/(2a)}\} = \sqrt{a^2 + b} = \sqrt{N}. \quad (6.7)$$

Brezinski (see books) reports that the solution was published by Landry in 1856. The first few convergents are

$$C_0 = a, \quad C_1 = \frac{2a^2 + b}{2a}, \quad C_2 = \frac{a(4a^2 + 3b)}{4a^2 + b}, \quad C_4 = \frac{8a^4 + 8a^2b + b^2}{4a(2a^2 + b)}$$

and the numerators and denominators follow the recursion relations

$$p_{k+1} = 2ap_k + bp_{k-1}, \quad q_{k+1} = 2aq_k + bq_{k-1}.$$

Evaluating a few of these shows the pattern

$$p_0^2 - Nq_0^2 = -b, \quad p_1^2 - Nq_1^2 = b^2, \quad p_2^2 - Nq_2^2 = -b^3.$$

In general

$$p_k^2 - Nq_k^2 = (-b)^{k+1}. \quad (6.8)$$

Therefore a solution set is $x = p_k$, $y = q_k$, $z = -b$, $n = k + 1$.

Example Find sets of integers (x, y, z, n) to satisfy $x^2 - 69y^2 = z^n$.

Solution : $N = 69 = 8^2 + 5$. The convergents of $\{8 : \underline{5/16}\}$ are 8, 133/16, 2168/261, 35353/4256, 576488/69401, 9400573/1131696, etc⁶. Therefore solutions are

$$133^2 - 69 \times 16^2 = 25 = 5^2, \quad 2168^2 - 69 \times 261^2 = -125 = -5^3, \quad 35353^2 - 69 \times 4256^2 = 625 = 5^4,$$

$$576488^2 - 69 \times 69401^2 = -3125 = -5^5, \quad 9400573^2 - 69 \times 1131696^2 = 15625 = 5^6, \quad \text{etc.}$$

As noted in §3.8, these representations in generalised continued fractions are not unique, and other ways of writing $\sqrt{69}$ are readily found; for example $\{7 : \underline{10/7}, \underline{5/7}\}$. When the convergents of this are substituted into $x^2 - 69y^2$, the sequence of values for $k = 1, 2, \dots$ has the interesting pattern

$$-2^25^1, \quad 2^25^2, \quad -5^3, \quad 2^25^4, \quad -2^25^5, \quad 5^6, \quad -2^25^7, \quad 2^25^8, \quad -5^9, \quad 2^25^{10}, \quad \text{etc.}$$

Therefore, in addition to $2168^2 - 69 \times 261^2 = -5^3$ from above, we have $224^2 - 69 \times 27 = -5^3$.

Another fanciful representation, similarly derived, is

$$\sqrt{69} = \{6 : \underline{\frac{33}{2}/6}, \underline{\frac{33}{4}/6}\}.$$

Be aware that the fraction in the numerator must not be factored into the denominator in evaluating the convergents. For increasing k the sequence of values of $p_k^2 - 69q_k^2$ is

$$-3.11^1, \quad 11^2, \quad -3.11^3, \quad 11^4, \quad -3.11^5, \quad 11^6, \quad \text{etc.}$$

Clearly one can play tunes on the generalised continued fractions to produce other interesting solutions to Eq 6.6.

⁶Bear in mind that convergents must always be calculated without any cancellation between numerator and denominator.

7 The CFRAC algorithm for factorising large integers

The CFRAC algorithm hit the headlines in 1970 when it was used by Morrison and Brillhart to factorise the 7th Fermat number, $2^{2^7} + 1$. Since it draws heavily on the best fit property of continued fractions, I describe it here. There are two parts to the theory behind this algorithm to factorise N :

- Legendre congruences of square integers x^2, y^2 to identify possible factors of N , and
- continued fraction expansion of \sqrt{N} to find suitable trial values of x and y .

7.1 Legendre congruence

We are given the composite integer N and wish to find all its prime factors. Suppose by some means we can find two positive integers $x < N, y < N$ whose squares have the same remainder when divided by N . That is

$$x^2 \equiv y^2 \pmod{N} . \quad (7.1)$$

Integers with this property are said to have Legendre congruence, after the distinguished French pioneer. Then

$$(x+y)(x-y) = x^2 - y^2 \equiv 0 \pmod{N} \quad \text{meaning that} \quad (x+y)(x-y) = \lambda N \quad (7.2)$$

for some integer λ , which could be 1. There are three possibilities:

1. Neither $x+y$ nor $x-y$ has a factor in common with N ; they both divide only λ .
2. Either $x+y$ or $x-y$ is equal to N , or to a multiple of N .
3. Either $x+y$ or $x-y$, or both, has at least one factor > 1 in common with N . This will be the $\gcd(x \pm y, N)$.

The first possibility corresponds to $\gcd(x \pm y, N) = 1$. It can yield no information on the factorisation of N . In case 2) N is cancelled from both sides on Eq 7.2. However case 3) holds the possibility that, by looking at $\gcd(x+y, N)$ and $\gcd(x-y, N)$, we may identify non-trivial factors of N . These g.c.d. can be found using Euclid's algorithm, §2.1. Some examples will illustrate the method.

Example 1: Factorise 55. Make a list of squares mod 55 until you find two with the same remainder. As it happens, the first square > 55 is $64 = 55 + 9 \equiv 3^2 \pmod{55}$, so take $x = 8, y = 3$. Then $x+y = 11$ and $x-y = 5$. Both of these are prime so we have found the prime factorisation $55 = 5 \times 11$.

Example 2: Factorise 2093. This is harder work because we have to test squares mod 2093 up to 57^2 before finding two that are equivalent. $57^2 \equiv 1156 = 34^2 \pmod{2093}$, so $x+y = 57+34 = 91$ and $x-y = 57-34 = 23$, and $23 \times 91 = 2093$, so $\lambda = 1$. 23 is a prime factor of 2093. The composite 91 can itself be factorised by Legendre congruence by observing that $10^2 = 100 \equiv 3^2 \pmod{91}$, making $10+3 = 13$ and $10-3 = 7$ the other two prime factors. We conclude that $2093 = 7.13.23$.

Example 3a: Factorise 77507. $\sqrt{77507} = 278.4$ so, starting at 279, systematically check $279^2, 280^2, 281^2$, etc. until you find one congruent to one of the squares in the range 1^2 to $278^2 \pmod{77507}$. Luckily, fairly soon we come to $306^2 \equiv 16139 = 127^2$. This immediately gives $306+127 = 433$ and $306-127 = 179$ as candidate factors of some multiple of 77507. In fact $179 \times 433 = 77507$ so again $\lambda = 1$. Moreover both 179 and 433 are prime so the factorisation is complete.

The larger is N , the longer the search will be for congruent squares. We could do with some method for homing in on Legendre congruences. This is where continued fractions come in.

7.2 Guided search using continued fraction expansion of \sqrt{N}

In Pell's equation, §5, the key concept is that $p_k^2 - Nq_k^2 = \kappa_k$ where p_k/q_k is the k th convergent to \sqrt{N} and κ_k is small, sometimes 1. Since in factorising N we are looking for $x^2 - y^2 = (x+y)(x-y) \equiv 0 \pmod{N}$, we could examine $p_k^2 - \kappa_k = Nq_k^2$ since both sides are $\equiv 0 \pmod{N}$. If κ_k were a perfect square, we could identify p_k with x and $\sqrt{\kappa_k}$ with y , and consider $x+y$, $x-y$ as candidate factors of some multiple of N . In this way the convergents suggest fruitful integers to test for Legendre congruence.

Example 3b: Factorise 77507 with search guided by continued fractions. The continued fraction for $\sqrt{77507}$ is $\{ 278 : 2, 2, 50, 4, 1, 1, 2, 1, 1, 4, 50, 2, 2, 556 \}$ and the convergents are $278, 557/2, 1392/5, 70157/252, \text{etc.}$ We could directly evaluate expressions like $1392^2 - 5^2 \times 77507 = -11$ to find κ_2 , but for higher convergents this would involve squaring some very large integers. It is better to calculate $p_k \pmod{N}$, then $\kappa_k = p_k^2 - Nq_k^2 \equiv p_k^2 \pmod{N}$. Following this approach, Table 13 lists p_k and $\kappa \pmod{N}$ for the first few convergents. κ has been factorised because we are searching for squares.

Table 13: Numerators p_k of convergents to $\sqrt{77507}$ and their squares, mod 77507.

k	a_k	p_k	$p_k^2 \equiv \kappa_k$
0	278	278	-223
1	2	557	221=13.17
2	2	1392	77496 \equiv -11
3	50	70157 \equiv -7350	121 = 11 ²
4	4	282020 \equiv 49499 \equiv -28008	77224 \equiv -283
5	1	352177 \equiv 42149	254 = 2.127
6	1	634197 \equiv 14141	-179
7	2	1620571 \equiv 70431 \equiv -7076	254 = 2.127
8	1	2254768 \equiv 7065	77224 \equiv -283
9	1	3875339 \equiv 77496 \equiv -11	121 = 11 ²
10	4	17756124 \equiv 7021	77496 \equiv -11
11	50	891681539 \equiv 41011 \equiv -36496	221
12	2	1801119202 \equiv 11536	-223
13	2	4493919943 \equiv 64083	1
14	556	2500420607510	-223

The row for C_3 gives $(-7350)^2 = (+7350)^2 \equiv 11^2 \pmod{77507}$. Taking $x = 7350$, $y = 11$ we get 7339 and 7361 as factors of some multiple of N . In fact $7399 \times 7361 = 697 \times 77507$; that is, $\lambda = 697$. By Euclid's method, $\gcd(77507, 7339) = 179$ and $\gcd(77507, 7361) = 433$, giving the prime factorisation of Example 3a).

A slightly different approach is to examine the prime factors of $\lambda = 679, 7399$ or 7361.

- Taking 697 as the easiest, use the CFRAC algorithm again or just division by small primes to find $697 = 17 \cdot 41$. So 17 must divide either 7399 or 7361, and so must 41. In fact $7361/17 = 433$ and $7339/41 = 179$.
- An alternative choice is to factorise 7361. Use the CFRAC algorithm. The C_3 convergent of $\sqrt{7361} = \{85: 1, 3, 1, 9, \dots\}$ is $429/5$ and gives $429^2 \equiv 4^2 \pmod{7361}$, from which the prime factor $433 = 429 + 4$ is found. Trial division then shows that $77507/433 = 179$.

In Table 13 note how the palindromic symmetry in the a_k carries into the κ_k . In particular, $\kappa_{13} = 1$, as would be expected from the analysis of Pell's equation in §5. The symmetry means that only convergents within one cycle of the recursion sequence, length L , can contribute independent information towards the factorisation process. However there is symmetry about the mid position at $k = L/2$ and this can reduce the number of useful convergents further. The symmetry depends on whether L is even or odd. If L is even, as it seems to be for about 90% of integers N , the κ_k values are mirrored about the $k = L/2$ position. If L is odd, the mirroring is accompanied by change of sign⁷.

Suppose that in the right column of Table 13 we did not find a value of p_k^2 which was a perfect square. It is valid to multiply rows for two or more convergents to form a square, so forming Legendre congruences by compounding rows. With short L , in *some* cases use can be made of the symmetry in the κ_k mentioned above. Here are two examples:

1. $p_1 = 557$ and $p_{11} = 41011 \equiv -36496$ both have $\kappa = 221$, being symmetrically placed about the 'mirror position' at p_7 . So $(557 \times -36496)^2 \equiv 221^2$. This gives $x = 557 \times 36496 \bmod N = 21438$, $y = 221$, $x + y = 21659$, $x - y = 21217$. The product $(x + y)(x - y) = 459539003 = 5929 \times 77507 = 77^2 N$, so $\lambda = 77^2$. $\gcd(x + y, 77507) = 179$ and $\gcd(x - y, 77507) = 433$.
2. Multiply the symmetrically placed p_4 and p_8 . If we take $x = 49499 \times 7065$ and $y = 283$, this makes $x + y = 349710718$ and 349710152 . $\gcd(x + y, 77507) = 433$, $\gcd(x - y, 77507) = 179$.

In both cases we recover $N = 77507 = 179 \times 433$. While this use of symmetry gives a factorisation in this case, in general it seems to depend on the particular N . Clearly, it is attractive to try this symmetry first, because it is very obvious which two convergents (rows in the table) to combine to form a square – they have the same κ .

Example 4: Factorise 46318663. This fourth example illustrates the effort and frustration involved in factorising a fairly large integer with CFRAC. The frustration comes because by no means all Legendre congruences lead to a factorisation of N – many are of category 1) or 2) of §6.1, with $x \pm y \equiv 0 \pmod N$. The method requires computer software which can manipulate arbitrarily large integers without loss of precision. Computer packages are freely available which use strings to accomplish this. I find that $L = 3296$ for \sqrt{N} . One *could* evaluate about 2000 convergents and compound those at positions k symmetrically placed about $k = 1648$ as in Example 3b. However, here we will follow Morrison and Brillhart and compound non-symmetrically placed rows from the first few dozen convergents.

First we need the continued fraction for $\sqrt{N} = \{6805: 1, 3, 1, 1, 2, 1, 2, 5, 2, 1, 1, 7 \dots\}$, $N = 46,318,663$. Table 14 lists the numerators $p_k \bmod N$ of the first 50 convergents. It also lists their squares $\kappa_k \bmod N$ and the factorisation of each square. Not one is a perfect square! Therefore our options are i) to determine yet more convergents in the hope of finding a square, or ii) to combine rows to form congruences from the p_k so far determined. Let's explore the latter.

We are looking for products of factorisations of κ_k which give a perfect square. Clearly this will need every prime factor to be raised to an even power. It is therefore helpful to organise the p_k^2 in order of highest factor, deleting from the list all which have a factor which occurs nowhere else in the list. For example, in Table 14 p_1^2 has 991 as a factor, but 991 does not occur elsewhere so p_1^2 cannot be a component of any square compounded from this table. When all the unpaired primes are removed (most of which are the larger primes), the ordered list looks like Table 15.

⁷ $N = 78259$ is an example of an integer in which the $x = p_k \bmod N$ values are also symmetrical about the $k = L/2$

Table 14: p_k and p_k^2 for $\sqrt{46318663}$, mod 46318663.

k	p_k	p_k^2	factors	k	p_k	p_k^2	factors
0	6805	-10638	-2.3 ³ .197	25	13283878	2546	2.19.67
1	6806	2973	3.991	26	43219939	-5039	-5039
2	27223	-6879	-3.2293	27	7086430	5538	2.3.13.71
3	34029	6266	2.13.241	28	11074136	-51	-3.17
4	61252	-4199	-13.17.19	29	34730837	6602	2.3301
5	156533	7362	2.3 ² .409	30	34217147	-767	-13.59
6	217785	-4687	-43.109	31	14279717	3321	3 ⁴ .41
7	592103	2362	2.1181	32	45017352	-559	-13.43
8	3178300	-5007	-3.1669	33	29366916	1401	3.467
9	6948703	5826	2.3.971	34	31407618	-8542	-2.4271
10	10127003	-7463	-17.439	35	14455871	4581	3 ² .509
11	17075706	1733	1733	36	14000697	-942	-2.3.157
12	37019619	-3438	-2.3 ² .191	37	25190977	1417	13.109
13	35497237	8909	59.151	38	9126175	-6891	-3.2297
14	26198193	-2451	-3.19.43	39	34317152	6618	2.3.1103
15	1334020	9277	9277	40	43443327	-1963	-13.151
16	27532213	-2886	-2.3.13.37	41	17065136	5226	2.3.13.67
17	37611996	9829	9829	42	31254936	-5079	-3.1693
18	18825546	-1531	-1531	43	33256345	2698	2.19.71
19	2941712	1317	3.439	44	25324327	-7599	-3.17.149
20	1924003	-2951	-13.227	45	12262009	5261	5261
21	10637724	6213	3.19.109	46	37586336	-7374	-2.3.1229
22	12561727	-7318	-2.3659	47	3529682	3373	3373
23	23199451	741	3.13.19	48	1856719	-4803	-3.1601
24	36403090	-10599	-3.3533	49	7243120	6113	6113

As a further aid to finding congruences, the information in Table 15 can be presented as a spreadsheet of powers of primes, as in Figure 5a. The primes are listed as column heading, the convergent index k labels the rows, and the numbers, (all 1) denote the power to which each particular prime is raised. By first taking a pair with the same highest prime factor, it is easy to see which other rows are also needed to give column sums which are all even numbers, corresponding to a square factor. This is illustrated in Figure 5b where p_{28} has been combined with the p_{10} , p_{19} , this pair having 439 as a factor. Question: why do some primes occur in the factorisations of κ_k for more than one k , while other primes of similar size seem not to occur at all?

About eight or nine congruences can be formed from this list, but unfortunately not one yields a factorisation of N . For instance, the combination in Figure 5b, from convergents $k = 10$, 19 and 28 gives $x = 46296274$, $y = -3.17.439 = 22389$, but $x + y = N$ and $\gcd(x - y, N) = 1$ so this combination is fruitless. Other combinations such as $\{13, 30, 40\}$, $\{4, 23, 28\}$, $\{4, 10, 19, 23\}$ and $\{14, 25, 32, 41\}$ all similarly give $x + y \equiv 0 \pmod N$. Other combinations fail because $x - y \equiv 0 \pmod N$:

position, with those of odd k index being reversed in sign. Another strange pattern in these continued fractions!

Table 15: p_k^2 ordered by highest prime factor for convergents of $\sqrt{46318663} \pmod{46318663}$.

k	p_k	p_k^2	factors
10	10127003	-7463	-17.439
19	2941712	1317	3.439
13	35497237	8909	59.151
40	43443327	-1963	-13.151
6	217785	-4687	-43.109
21	10637724	6213	3.19.109
37	25190977	1417	13.109
27	7086430	5538	2.3.13.71
43	33256345	2698	2.19.71
25	13283878	2546	2.19.67
41	17065136	5226	2.3.13.67
30	34217147	-767	-13.59
14	26198193	-2451	-3.19.43
32	45017352	-559	-13.43
4	61252	-4199	-13.17.19
23	23199451	741	3.13.19
28	11074136	-51	-3.17

$\{6, 14, 21\}$, $\{23, 25, 41\}$, $\{21, 23, 37\}$ and $\{6, 14, 23, 27\}$ are four examples.

By this stage quite a lot of effort has been expended to no avail. We have no option but to evaluate more convergents and sort through them for pairs. Table 16 gives the results. Here, at $k = 57$ we have a perfect square, but it too is useless because $x + y \equiv 0 \pmod{N}$. However, C_{58} and C_{62} give exactly the same factors in p_k^2 , and for these we do find a Legendre congruence which can unlock the factorisation of N . $x = 46087055$, $y = 2.3.17.97 = 9894$, $x^2 \equiv y^2 \pmod{N}$ and

$$\frac{(x+y)(x-y)}{N} = \frac{46096949 \times 46077161}{46318663} = \lambda = 45856603.$$

Using Euclid's algorithm, $\gcd(x+y, N) = 6521$ and $\gcd(x-y, N) = 7103$, and

$$N = 6521 \times 7103.$$

These are primes so this is our hard won result. Incidentally, on dividing through by N , we obtain the bonus result that $\lambda = 45856603 = 6487 \times 7069 = 13 \times 499 \times 7069$.

	-1	2	3	5	7	11	13	17	19	37	43	59	67	71	109	151	439
p10	1							1									1
p19			1														1
p13												1				1	
p40	1						1									1	
p6	1										1				1		
p21			1						1						1		
p37							1								1		
p27		1	1				1							1			
p43		1							1					1			
p25		1							1								
p41		1	1				1						1	1			
p30	1						1					1					
p14	1		1						1		1						
p32	1						1				1						
p4	1						1	1	1								
p23			1				1		1								
p28	1		1					1									

(a)

	-1	2	3	5	7	11	13	17	19	37	43	59	67	71	109	151	439
p10	1							1									1
p19			1														1
p28	1		1					1									

(b)

Figure 5: Search for congruences from continued fraction of $\sqrt{46318663}$.
Array *a*) lists powers of prime factors of the squares of numerators p_k of convergents.
b) shows one selection where sum of powers is even for each prime factor.

Table 16: p_k^2 for higher convergents of $\sqrt{46318663} \pmod{46318663}$.

k	p_k	p_k^2	factors	k	p_k	p_k^2	factors
50	9099839	-6926	-2.3463	60	19979671	-10062	-2.3 ² .13.43
51	16342959	3177	3 ² .353	61	42684758	2377	2377
52	11810053	-7919	-7919	62	5444051	-9894	-2.3.17.97
53	28153012	4458	2.3.743	63	1810146	2033	19.107
54	21797414	-3363	-3.19.59	64	14494781	-10259	10259
55	907928	7773	3.2591	65	16304927	1866	2.3.311
56	22705342	-4574	-2.2287	66	19687017	-5639	-5639
57	46318612	2601	3 ² .17 ²	67	9360298	3242	2.1621
58	22705138	-9894	-2.3.17.97	68	1449248	-8231	-8231
59	22705087	1781	13.137	69	10809546	3873	3.1291

7.3 Implementation of CFRAC on a computer

One might well ask whether there is any easy way of picking the sets of convergents which will lead to a non-trivial factorisation – picking winners from losers. I do not know of one. CFRAC clearly requires many convergents to be tabulated, sorted and combined, and so calls for implementation on a fast computer. Indeed it was the wider use of computers in the 1960s which led to development of CFRAC from earlier ideas. The book by Hans Riesel and the original article by Morrison and Brillhart explain how the CFRAC algorithm is implemented on computer. This includes methods for optimising the procedure and working with very large integers. However, I can offer a few first hand observations, based on the experience of writing my own unsophisticated program to implement CFRAC on a home computer.

The prerequisite is a programming language or set of library functions to carry out arithmetic to arbitrary precision. The stages in the program are as follows:

1. Decide how many convergents of \sqrt{N} you wish to use and expand \sqrt{N} as a continued fraction using the algorithm of §3.3. I expanded up to the recursion length L , or to $k = 200$ if $L > 200$.
2. Calculate a table of $p_k \bmod N$ and $\kappa_k \bmod N$.
3. Have prepared a data table of all primes up to some limit. I chose the first 500 primes, to 3571. Factorise each κ_k by dividing by each of these primes in turn, and store the index of each prime in an array, equivalent to Figure 5a). Also store the sign, + or –, of κ_k . If there is a residual quotient > 1 after all tabulated primes have been divided, store it too for the moment. (This residual will be larger than the largest prime in the data table, and we choose not to pursue factorising it.) At this stage you have a large array of rows labelled by convergent k and columns labelled by primes, each cell containing the corresponding power of that prime as a factor of κ_k .
4. Search each row to see if it corresponds to κ_k being a square, for which all indices $\equiv 0 \pmod 2$. If so $\kappa_k = y^2$. Evaluate $x + y$, $x - y$ and the two gcd. Some factors may be revealed at this stage, though possibly not a complete factorisation.
5. Reduce the array by deleting valueless rows and columns. Thus,
 - a) search the column of residual quotients to check whether any two happen to have the same value, but otherwise delete that column and all rows with residual values,
 - b) delete each column where all indices are zero, meaning that prime did not divide any κ_k ,
 - c) delete each column whose sole entry is a single 1, meaning that that prime divided only one κ_k . Delete also the row in which this 1 occurs. It cannot be paired with any other row to form a Legendre congruence.
6. Sort the array and search for combinations of rows such that the sum of indices in each column is even. This is the most tricky part of the algorithm and, for my purposes, I have found it convenient to carry out by hand on a spreadsheet. I therefore simply printed out and copied the reduced row and copy to a spreadsheet. Sort the rows in descending order of prime factor. Adjacent sorted rows should show paired large primes. Starting with the largest primes, build congruences by selecting lower rows with smaller primes, until the sum of indices is even for every prime, and the sign is +1. You then have a congruence. Note the row numbers k which make up this combination.
7. Return to the program and key in the row numbers for that congruence. Hence determine $x \bmod N$, $y^2 \bmod N$, $x + y$, $x - y$, $\gcd(x + y, N)$, $\gcd(x - y, N)$.

8. Repeat with other congruences until you have a complete factorisation.

When the program is run for $N = 463189663$, Example 4, it gives κ_{131} and κ_{183} as squares, and both immediately give the factorisation 6521×7103 . There is no need to make further combinations of rows using the spreadsheet, though an example of a successful combination is $\{k = 23, 76, 98\}$. Riesel, and Morrison and Brillhart, explain how in professional implementations of CFRAC, matrix algebra techniques akin to Gaussian elimination are applied within the one program to carry out the sorting and pairing functions which I have done by hand using the spreadsheet, but the outcome is the same.

7.4 Factorisation of integers $N = a^2 + 1$

The power of CFRAC relies on obtaining a sufficiently long table of convergents of \sqrt{N} that Legendre congruences can be formed. Some integers, however, are ‘awkward’ because the continued fraction expansion of \sqrt{N} has a very short recursion sequence which does not furnish enough suitable convergents. §3.2 and 3.4 discuss several types with recursion lengths $L = 1, 2$ and 3 . The extreme type is $L = 1$ which we know from §3.2 corresponds to $N = a^2 + 1$ for some integer a . For this $\sqrt{N} = \{a : \underline{2a}\}$. Convergent C_1 is

$$\frac{2a^2 + 1}{2a} \quad \text{with numerator } p_1 = (a^2 + 1) + a^2 \equiv a^2 \equiv -1 \pmod{N}$$

so $p_k^2 \equiv +1$. Continuing, C_3 is

$$\frac{8a^2(a^2 + 1) + 1}{4a(2a^2 + 1)} \quad \text{with numerator } p_2 \equiv +1 \pmod{N},$$

making $\kappa_3 \equiv 1 \pmod{N}$ also. Indeed, p_k alternates between $+1$ and -1 indefinitely, which is useless for Legendre congruences.

Fortunately there is an easy way out of this problem, provided by the absence of an obvious link between the continued fraction of N and that of any of its multiples; recall the example of multiples of $7/23$ in §1.4. We simply try various multiples of N , picking out those which give a long enough recursion sequence to furnish a useful set of congruences.

Example 5: Factorise 901. $901 = 30^2 + 1$ so we have to take multiples. The recursion lengths L of $2 \times, 3 \times, 4 \times, 5 \times, 6 \times,$ and 7×901 are respectively $4, 2, 2, 9, 26$ and 4 , so both $N' = 5 \times 901 = 4505 = \{67 : 8, 2, 1, 1, 1, 1, 2, 8, 134\}$ and $N'' = 6 \times 901$ have potential. Table 17 gives results for 4505 . Even these few convergents seem full of possibilities:

- κ_1 is a square and gives $x + y = 544, x - y = 530, \gcd(x + y, 4505) = 17, \gcd(x - y, 4505) = 265 = 3.53$.
- κ_5 is another square but has $\gcd(x + y, N) = N, \gcd(x - y, N) = 1$ so yields no information.
- κ_7 is yet another square and gives $x + y = -1105, x - y = -1113, \gcd(x + y, N) = 85 = 17.5, \gcd(x - y, N) = 53$.
- A larger combination is rows 0, 3 and 4. For this $x = 67.1678.2819 \equiv 2144, y = 4.59 = 236, x + y = 2380, x - y = 1908, (x + y)(x - y)/4055 = 1008 = \lambda, \gcd(x + y, N') = 85 = 5.17, \gcd(x - y, N') = 53$.

Table 17: p_k^2 for convergents of $\sqrt{5 \times 901} \pmod{4505}$.

k	p_k	$p_k^2 = \kappa_k$	factors
0	67	-16	-4^2
1	537	49	7^2
2	1141	-64	-8^2
3	1678	59	59
4	$2819 \equiv -1678$	-59	-59
5	$4497 \equiv -8$	64	8^2
6	$7316 \equiv 2811$	-49	-7^2
7	1109	16	4^2

All these sets of convergents lead to $5 \times 901 = 5 \times 17 \times 53$, making $901 = 17 \times 53$.

This route to factorising $N = a^2 + 1$ is essential for the celebrated Fermat numbers of the form $2^{2^n} + 1$. When Morrison and Brillhart factorised the 7th Fermat number, $2^{128} + 1$, they used the multiplier 257. They determined that $F_7 = 340282366920938463463374607431768211457 = 59649589127497217 \times 5704689200685129054721$. CFRAC was the first significant computer-based method for factorisation, though nowadays it has been superseded largely by the quadratic sieve and number field sieve methods. Let us put all this together by attempting the factorisation of $F_5 = 2^{2^5} + 1$. This was achieved by the blind Euler in 1732, so we should be able to achieve it in 2011!

Example 6: Factorise $F_5 = 2^{32} + 1 = 4294967297$. It has the form $N = a^2 + 1$ so we need the continued fraction of a multiple, \sqrt{mN} . This poses some questions of strategy:

1. Since obtaining a single κ_k which is a square is far simpler than compounding several rows in search of a congruence, do we focus on finding a multiplier m which produces such a single square κ within, say 200 or 500 convergents? That is, if we fail to find a square κ within that number of convergents, do we abandon that m and try another, rather than spend effort on compounding rows in search of a congruence?
2. Is there any way of knowing in advance what the most fruitful multiple m will be?
3. For a given m and N , what is the most efficient number of convergents to search for a square. We might expect that the larger mN , the higher the C_k , κ_k which must be evaluated.

I have looked at a few multiples of F_5 to see which give a square within the first 500 convergents.

1. $2F_5$ has $\kappa_{183} \equiv 201^2$, giving $\gcd(x + y, 2F_5) = 2F_5$, $\gcd(x - y, 2F_5) = 2$, which is true but unhelpful. However $\kappa_{339} \equiv 103^2$ gives $\gcd(x + y, 2F_5) = 1282 = 2 \times 641$, $\gcd(x - y, 2F_5) = 13400834 = 2 \times 6700417$.
2. $3F_5$ has $\kappa_{191} \equiv 29^2$, giving $\gcd(x + y, 3F_5) = 641$, $\gcd(x - y, 3F_5) = 20101251 = 3 \times 6700417$. Also $\kappa_{283} \equiv 71^2$, giving $\gcd(x + y, 3F_5) = 6700417$, $\gcd(x - y, 3F_5) = 1923 = 3 \times 641$.
3. $5F_5$ has $\kappa_{267} \equiv 366^2$, and $\kappa_{297} \equiv 77^2$. Both these have $\gcd(x + y, 5F_5) = 6700417$, $\gcd(x - y, 5F_5) = 3205 = 5 \times 641$.

4. $7F_5$ has $\kappa_{229} \equiv 29^2$, giving $\gcd(x + y, 7F_5) = 46902919 = 7 \times 6700417$, $\gcd(x - y, 7F_5) = 641$.

5. $17F_5$ has no squares in the first 500 values of κ_k .

6. $19F_5$ has $\kappa_{175} \equiv 191^2$, giving $\gcd(x + y, 19F_5) = 6700417$, $\gcd(x - y, 19F_5) = 19 \times 641$.

If you were to limit yourself to only 200 convergents, $5F_5$ would not have shown a useful square κ , and you would have to combine rows to form a congruence. One fruitful compound congruence is from the four rows, $\{k = 87, 99, 127, 189\}$. For this

$$x \bmod 5F_5 = -17689213053, \quad y^2 = 775072307338671093444, \quad y \bmod 5F_5 = 6365283977$$

$$x + y = -11323929076, \quad \gcd(x + y, N) = 641,$$

$$x - y = -24054497030, \quad \gcd(x - y, N) = 33502085 = 2 \times 6700417.$$

Clearly, from all this evidence, the factorisation of $F_5 = 641 \times 6700417$. Note that there is no evidence of any factors other than these two, except trivially for the multiplier m . This points strongly to 6700417 being prime, which indeed it is. So CFRAC has a secondary application as a method for testing for primes.

I cannot resist closing this section on CFRAC without pointing out a charming and elegant proof that 641 is a factor of F_5 using modular arithmetic, without direct trial division. It is based on two partitions of 641:

$$641 = 5 \cdot 128 + 1 = 5 \cdot 2^7 + 1, \quad 641 = 16 + 625 = 2^4 + 5^4$$

$$\text{So } 5 \cdot 2^7 \equiv -1 \pmod{641}, \quad \text{and } 5^4 \equiv -2^4 \pmod{641}.$$

$$(5 \cdot 2^7)^4 = 5^4 \cdot 2^{28} \equiv +1 \pmod{641}$$

$$-2^4 \cdot 2^{28} = -2^{32} \equiv 1 \pmod{641}$$

$$2^{32} + 1 \equiv 0 \pmod{641}.$$

Lovely! Incidentally, Euler proved that every Fermat number F_n that factorises must have a prime factor of the form $c \cdot 2^n + 1$. In his book, Hans Riesel tabulates dozens of these.

Part III

Convergents: Approximation to Reals

This Part extends the analysis of properties of the sequence of convergents, C_k , and their errors, ϵ_k , which were introduced in Part I. §8 discusses interpretation of the recursion relations Eq 1.2 and 1.3 as a functional mapping (*i.e.* a transformation) and its representation by matrices formed from the convergents. These mappings and matrices are related to certain infinite mathematical symmetry groups. §9 gives proofs of the ‘best fit’ property and the ‘ $1/2v^2$ test’ for whether a given fraction, u/v , is a convergent of a given real θ . Subsequent sections examine the rate of convergence of the C_k sequence, and the use of ϵ_k as a criterion for distinguishing various categories of real number, from rationals, to quadratic surds, to higher algebraic irrationals and then to the transcendental numbers.

8 Groups and matrices associated with convergents

Just to remind the reader, we will be building on the crucial relations Eqs 1.1 to 1.6:

$$\theta \equiv \theta_0 = a_0 + \rho_0, \quad \frac{1}{\rho_0} = \theta_1 = a_1 + \rho_1, \quad \frac{1}{\rho_1} = \theta_2 = a_2 + \rho_2, \quad \text{etc.} \quad \text{Copy of (1.1)}$$

$$p_{k+1} = a_{k+1}p_k + p_{k-1} \quad \text{and} \quad q_{k+1} = a_{k+1}q_k + q_{k-1}. \quad \text{Copy of (1.2)}$$

$$\theta \equiv \theta_0 = \frac{\theta_{k+1}p_k + p_{k-1}}{\theta_{k+1}q_k + q_{k-1}}. \quad \text{Copy of (1.3)}$$

$$p_k q_{k-1} - p_{k-1} q_k = (-1)^{k-1} \quad \text{(Copy of 1.5)}$$

$$\Delta_k \equiv C_k - C_{k-1} = \frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} = \frac{(-1)^{k-1}}{q_k q_{k-1}} \quad \text{(Copy of 1.6)}$$

In Eq 1.3 $\theta_{k+1} > 1$ (strictly > 1) is the next complete quotient. Eq 1.3 states that θ can be expressed in terms of any pair of adjacent convergents together with the next complete quotient.

Both Eq 1.2 and 1.3 have the same form. In them two numbers, θ and ξ , say, are related by a linear rational function

$$\theta = \mathcal{T}(\xi) = \frac{\xi A + C}{\xi B + D} \quad (8.1)$$

where A, B, C, D are constants. The notation $\mathcal{T}(\xi)$ denotes a transformation operator acting on ξ to produce θ . It implies a geometric interpretation in which a point ξ in the real number line is mapped to another point θ . So we distinguish between

- i) the transformation operation and
- ii) the set on which the transformation operates.

8.1 Background on bilinear mapping functions

Focusing first on the transformation operation, there is a wide class of mapping functions with the form Eq 8.1. In the most general case A, B, C, D are complex constants, and ξ and θ complex variables. They have a geometrical interpretation as transformations in the complex plane \mathbb{C} . With Eq 8.1 is associated a matrix representation of the numerator and denominator:

$$\begin{pmatrix} A & C \\ B & D \end{pmatrix} \begin{pmatrix} \xi \\ 1 \end{pmatrix} = \begin{pmatrix} \xi A + C \\ \xi B + D \end{pmatrix}$$

If the determinant $AD - BC$ is non-zero, the system can be inverted and gives

$$\xi = \mathcal{T}^{-1}(\theta) = \frac{\theta D - C}{-\theta B + A} \quad (8.2)$$

where $\mathcal{T}^{-1}(\theta)$ is the inverse operator of $\mathcal{T}(\xi)$. It has the same linear rational form. Such invertible transformations are known as linear fractional, bilinear or Möbius transformations after the German geometer August Möbius who first published their properties and application to projective geometry, in about 1830. Möbius transformations map circles to circles in the complex plane and on the Riemann sphere. They also preserve angles and orientation, and include as special cases dilations, translations, rotations and inversions through a point (but not reflections, which reverse orientation). They form an infinite symmetry group, the Möbius group, under the operation of Eq 8.1, 8.2, as described in textbooks on projective and hyperbolic geometry.

There is an important subgroup of the Möbius group in which the constants are restricted as follows:

- A, B, C, D are integers,
- the determinant $AD - BC = +1$ or -1 .

A matrix with these restrictions is called *unimodular*. This subgroup thus represented is the positive/negative special linear group $S^\pm L(2, \mathbb{Z})$. It too can act on all numbers in the complex plane, \mathbb{C} . The significance of the determinant having modulus 1 is that its matrix is in a sense normalised to unit size because, viewed geometrically, the determinant of $\begin{pmatrix} A & C \\ B & D \end{pmatrix}$ is the area of the parallelogram spanned by the vectors (A, B) and (C, D) . There is an important subgroup of $S^\pm L(2, \mathbb{Z})$ in which the determinant is strictly $+1$. This is the modular group, which is isomorphic to the projective special linear group $\text{PSL}(2, \mathbb{Z})$.

Restriction of the constants to integers means that the group acts discretely on the complex plane, segmenting it into a type of lattice. Discrete subgroups of the Möbius group were studied extensively by the Polish/German mathematician Lazarus Fuchs in connection with differential equations. They were developed further by the eminent French mathematician Henri Poincaré, who named them Fuchsian groups in Fuchs' honour. They are discussed at length in books on, for instance, hyperbolic geometry.

The numerators p_k, p_{k-1} and denominators q_k, q_{k-1} of adjacent convergents of $\theta \in \mathbb{R}$ are integers and, from Eq 1.5, have a cross product equal to ± 1 . Building on the matrix notation of §1.4.1 of Part I, we can therefore associate a matrix of the above Möbius form with these consecutive convergents

$$\mathcal{M} = \begin{pmatrix} p_k & p_{k-1} \\ q_k & q_{k-1} \end{pmatrix}. \quad (8.3)$$

This matrix represents one element of the group $S^\pm L(2, \mathbb{Z})$. Four points to note are:

- With convergents (as opposed to general positive integers) $p_k > p_{k-1}, q_k > q_{k-1}$; that is, the first column of the matrix always exceeds the second.
- Swapping rows gives the matrix associated with the reciprocal fractions $q_k/p_k, q_{k-1}/p_{k-1}$. This matrix has the opposite sign of determinant, consistent with the reciprocal continued fraction being shifted by one place, so that $k \rightarrow k \pm 1$ and $(-1)^{k-1}$ changes sign in Eq 1.5.

- For two matrices M_1 and M_2 , $\det(M_1 M_2) = \det(M_1) \cdot \det(M_2)$, so all products and positive powers of matrices of convergents like Eq 8.3 are also elements within $S^\pm L(2, \mathbb{Z})$. In group theory terms, repeated multiplication by the same matrix (with $p_k, p_{k-1}, q_k, q_{k-1}$ all fixed) steps an element $\begin{pmatrix} \xi \\ 1 \end{pmatrix}$ along a trajectory, part of an orbit.
- The inverse of matrix Eq 8.3 is

$$\mathcal{M}^{-1} = \begin{pmatrix} q_{k-1} & -p_{k-1} \\ -q_k & p_k \end{pmatrix} \quad (8.4)$$

and is also an element of $S^\pm L(2, \mathbb{Z})$. However, having negative signs, it cannot correspond to partial quotients. This proves that matrices associated only with convergents cannot form a subgroup of $S^\pm L(2, \mathbb{Z})$. If the entries in Eq 8.4 are replaced by their absolute values, the first column corresponds to the ratio of adjacent denominators which, from §2.3, Part I, has the same partial quotients as p_k/q_k but in reverse order.

8.2 Farey sequences: $ad - bc = \pm 1$ implies ‘no intermediate u/v ’

There is also a significant similarity of discrete Möbius functions to fractions in a Farey sequence. These are named after John Farey, an English geologist and writer, who published a note on them in 1816. The Farey sequence F_n of order n is the sequence of fractions a/b between 0 and 1, expressed in lowest terms and placed in order of increasing size, which have denominators $1 \leq b \leq n$. Thus

$$F_4 = \frac{0}{1}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{1}{1}.$$

$$F_7 = \frac{0}{1}, \frac{1}{7}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{2}{7}, \frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \frac{1}{2}, \frac{4}{7}, \frac{3}{5}, \frac{2}{3}, \frac{5}{7}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \frac{1}{1}.$$

The similarity with the Möbius functions is that if $a/b, c/d$ are a neighbouring pair in a Farey sequence, then $ad - bc = \pm 1$. Conversely, if $ad - bc = \pm 1$ for positive integers a, b, c, d , then $a/b, c/d$ are neighbours in the Farey sequence F_n where $n = \max(b, d)$. The sign of $ad - bc$ is positive if $a/b > c/d$.

Looking further into this, any two consecutive numbers, when factorised, can provide two fractions, a/b and c/d , such that $ad - bc = \pm 1$. As numerical examples, consider Table 18, where each column is created from consecutives. In all cases $ad - bc = -1$ but the bottom row shows that in some cases $b < d$, in other $b > d$, depending on how the integers factorise. Each of these pairs must be consecutive fractions in a Farey sequence.

consecutives	32, 33	33, 34	34, 35	35, 36	36, 37	37, 38	38, 39	39, 40
fraction pair	$\frac{4}{11} < \frac{3}{8}$	$\frac{3}{17} < \frac{2}{11}$	$\frac{2}{7} < \frac{5}{17}$	$\frac{5}{12} < \frac{3}{7}$	$\frac{6}{37} < \frac{1}{6}$	$\frac{1}{19} < \frac{2}{37}$	$\frac{2}{3} < \frac{13}{19}$	$\frac{3}{5} < \frac{8}{13}$
$b > d?$	y	y	n	y	y	n	n	n

Table 18: Paired fractions $\frac{a}{b} < \frac{c}{d}$, with $ad - bc = -1$, created from consecutive integers.

An important property of such fraction pairs is that there is no fraction u/v with a value between a/b and c/d which also has a denominator between b and d . Roughly speaking, if $b < v < d$ or $d < v < b$, the constraint $ad - bc = \pm 1$ brings a/b and c/d too close together for any u/v to squeeze in between. Contrast this with the fractions $1/5$ and $1/11$, say, where $1/6, 1/7, \dots, 1/10$ all lie between. Here $ad - bc = 11 - 5 = 6$. Now change the second fraction to $2/11$, making $ad - bc = 11 - 10 = 1$ and there is no intermediate fraction with denominator 6, 7, 8, 9, or 10. The relevant section of the F_{11} Farey sequence is $\dots, \frac{1}{6}, \frac{2}{11}, \frac{1}{5}, \frac{2}{9}, \dots$. This ‘no intermediate u/v ’ property when $ad - bc = \pm 1$ is closely related to the best fit property of continued fractions, stated in Eq 1.11 and proved in §9,

that any fraction u/v which is a closer approximation to a given θ than its convergent p_k/q_k must have a larger denominator; that is $v > q_k$.

In proving the ‘no intermediate u/v ’ property there are two cases to consider: $b < v < d$ and $d < v < b$. In the first case put the three fractions over a common denominator and use $bc = ad + 1$:

$$\frac{a}{b} < \frac{u}{v} < \frac{c}{d} \implies adv < bdu < adv + v \implies 0 < d(bu - av) < v$$

Bearing in mind that $bu - av \neq 0$ is an integer, when $d > v$ there can be no multiple of d less than v , so no such u/v exists. To deal with the second case of $b > v$, replace ad by $bc - 1$:

$$\frac{a}{b} < \frac{u}{v} < \frac{c}{d} \implies (bc - 1)v < bdu < bcv \implies 0 < b(du - cv) + v < v$$

and now no integer multiple of b can make this true. With both cases excluded, the proof is complete.

One further property of Farey series will be used in §9.3. It concerns the conditions for three fractions $a/b < c/d < f/g$ to be consecutive in a Farey series. Two simultaneous equations are formed from $ad - bc = cg - df = -1$ with solution

$$c = \frac{d(a+f)}{b+g}, \quad d = \frac{-(b+g)}{ag-bf} \quad \text{from which} \quad \frac{c}{d} = \frac{a+f}{b+g}.$$

This pleasing result says that the intermediate fraction is formed by adding the numerators and denominators of the fractions immediately below and above. §9.3 uses this in the form that, for $r < p, s < q$,

$$\frac{r}{s}, \quad \frac{p}{q}, \quad \frac{p+r}{q+s} \tag{8.5}$$

are three consecutive Farey fractions in increasing order.

8.3 Equivalent numbers

We now move the focus from the Möbius matrix transformations to the sets on which they operate. In their classic book (page 141) Hardy and Wright discuss ‘equivalent numbers’ in relation of convergents of continued fractions. Since, the name ‘equivalent numbers’ is used in several different senses in maths, here they must be understood as linear fractional equivalent numbers, as defined below. They are a generalisation of the usual equivalent fractions $\in \mathbb{Q}$ which allows all reals $\in \mathbb{R}$ to be separated into equivalence classes.

Hardy and Wright define two real numbers, ξ, θ , as being equivalent if the one can be obtained from the other by action of an element of the group $S^\pm L(2, \mathbb{Z})$. The definition is restricted to reals, in contrast with wider action of $S^\pm L(2, \mathbb{Z})$ on the \mathbb{C} plane. To be precise,

$\xi, \theta \in \mathbb{R}$ are ‘equivalent’ if there exist integers A, B, C, D with $AD - BC = \pm 1$ such that ξ maps to θ according to

$$\theta = \mathcal{T}(\xi) = \frac{\xi A + C}{\xi B + D}. \tag{8.1} \text{ Copy of}$$

The group nature of $S^\pm L(2, \mathbb{Z})$ ensures that the set of reals satisfying the above definition form an equivalence class. Thus, the identity matrix ensures that the relation is reflexive, the existence of a unique inverse, Eq 8.2, ensures a symmetrical relation, and the composition of operations by matrix multiplication ensures transitivity.

Hardy and Wright prove (page 144) that, in terms of continued fractions, ξ and θ are equivalent only if they have a complete quotient in common:

$$\xi = \{a_0 : a_1, a_2, \dots, a_{k-1}, \xi_k\}, \quad \theta = \{b_0 : b_1, b_2, \dots, b_{j-1}, \theta_j\}$$

where $\xi_k = \theta_j$. It will be clear that both are equivalent to the tail of their respective partial quotient sequence because, by Eq 1.3,

$$\xi = \frac{\xi_k p_{k-1} + p_{k-2}}{\xi_k q_{k-1} + q_{k-2}}, \quad p_{k-1} q_{k-2} - p_{k-2} q_{k-1} = \pm 1,$$

and similarly for θ . By the transient property of an equivalence relation, $\xi \equiv \xi_k = \theta_j \equiv \theta$ implies that $\xi \equiv \theta$. So equivalence can be expressed in terms of continued fractions by saying that ξ and θ are equivalent if they share the same infinite tail in their sequences of partial quotients.

Important for our purposes are the classes of equivalent numbers generated by $S^\pm L(2, \mathbb{Z})$ transformation of a square root, \sqrt{N} , N an integer: that is, numbers of the form

$$\frac{A\sqrt{N} + C}{B\sqrt{N} + D} = \frac{(ABN - CD) - \sqrt{N}}{B^2N - D^2}, \quad AD - BC = \pm 1.$$

This equivalence class lies in the field extension $\mathbb{Q}(\sqrt{N})$. However, not all numbers in this field extension are equivalent to \sqrt{N} as an example will make clear. Consider the Golden Mean $G = \frac{1}{2}(1 + \sqrt{5}) = \{1 : \underline{1}\}$. This is equivalent to $\frac{1}{2}(7 - \sqrt{5}) = \{2 : 2, \underline{1}\}$ through multiplication of $\begin{pmatrix} 1+\sqrt{5} \\ 2 \end{pmatrix}$ by $\begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix}$, determinant -1 . However, G is not equivalent to $\frac{1}{2}\sqrt{5} = \{1 : \underline{8}, \underline{2}\}$ even though it can be obtained by multiplication by $\begin{pmatrix} 1 & 2 \\ 2 & 0 \end{pmatrix}$, because this has determinant -4 .

An important particular case is that all rational numbers \mathbb{Q} are equivalent. This is because all can be expressed with the last complete quotient equal to 1 (if it is m , rewrite it as $[m-1] + 1/1$).

Some writers find it necessary to distinguish two varieties of equivalence depending on whether the determinant of the matrices which transform θ to ξ and *vice versa* is $+1$ or -1 . Those related by a $+1$ matrix are said to be ‘properly equivalent’ and those with -1 are ‘improperly equivalent’. This distinction will be referred to again in §11.2 in the context of quadratic forms.

8.4 Powers of a matrix and periodic continued fractions

First a reminder about orbits in group theory. When element g_1 of a group G acts on an element x of a set X , g_1 maps x onto another element $x' = g_1(x)$. Now repeat this with every element g_j of G . The ‘orbit of x under G ’ is the subset of X which consists of all elements $g_j(x)$, $j = 1, 2, \dots$. The orbit of x is the set of elements ‘reached’ from x by action of any and all elements of G .

Here, however, I wish to describe a subset of an orbit, restricted to the ‘trajectory’ of one point under the repeated action of a single given matrix in $S^\pm L(2, \mathbb{Z})$. Form the transformation matrix \mathcal{M} as in Eq 8.3 from two adjacent convergents of some real θ . In our analysis it will be important that p_{k-1}/q_{k-1} is an authentic convergent⁸ of p_k/q_k as well as being a convergent of θ – a detail explained below. Let \mathcal{M} act repeatedly on the ‘point’ $\mathcal{J} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ to generate a sequence of column matrices $\mathcal{M}^N \mathcal{J}$, each representing a fraction. Thus for $\mathcal{N} = 2$,

$$\begin{pmatrix} p_k & p_{k-1} \\ q_k & q_{k-1} \end{pmatrix}^2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} p_k^2 + p_{k-1}q_k & p_{k-1}(p_k + q_{k-1}) \\ (p_k + q_{k-1})q_k & p_{k-1}q_k + q_{k-1}^2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} p_k^2 + p_{k-1}q_k \\ (p_k + q_{k-1})q_k \end{pmatrix}.$$

⁸The distinction between an authentic and a quasi-convergent was made in §2.3, Part I.

The common fraction represented by the resulting column matrix will have a continued fraction expansion, and you might expect that multiplying repeatedly by the same matrix will have a recursive effect on this continued fraction. In fact each multiplication inserts a fixed sequence of partial quotients into the continued fraction. This is best illustrated numerically. Consider the powers of $\mathcal{M}_1 = \begin{pmatrix} 9 & 1 \\ 37 & 4 \end{pmatrix}$ acting on \mathcal{J} . $\mathcal{M}_1^1 \mathcal{J}$ represents the fraction $\frac{9}{37} = \{0 : 4, 9\}$. $\mathcal{M}_1^2 = \begin{pmatrix} 118 & 13 \\ 481 & 53 \end{pmatrix}$, so $\mathcal{M}_1^2 \mathcal{J}$ represents $\frac{118}{481} = \{0 : 4, 13, 9\}$. Note the insertion of $a_2 = 13$. Observe these higher powers:

- $\mathcal{M}_1^3 \mathcal{J}$ represents $\frac{1543}{6290} = \{0 : 4, 13, 13, 9\}$.
- $\mathcal{M}_1^4 \mathcal{J}$ represents $\frac{20177}{82251} = \{0 : 4, 13, 13, 13, 9\}$.
- $\mathcal{M}_1^8 \mathcal{J}$ represents $\frac{589955302}{2404936989} = \{0 : 4, 13, 13, 13, 13, 13, 13, 13, 9\}$.

Clearly, as the exponent $\mathcal{N} \rightarrow \infty$, the continued fraction tends to $\theta = \{0 : 4, \underline{13}\}$. This must be a quadratic surd. Using the methods of §3.2.2, Part I:

$$\theta = \frac{1}{4 + \eta} \quad \text{where} \quad \eta = \frac{1}{13 + \eta} \quad \text{from which} \quad \theta = \frac{1}{74}(5 + \sqrt{173}) = 0.24531.$$

Note that $173 = 13^2 + 4$ (*c.f.* §3.2.1). This gives some insight into how a finite continued fraction makes the transition to an infinite one, and so complements §3.1.

There are several subtleties associated with this. One is the effect of the two alternative ways of writing the last partial quotient in a finite continued fraction. §2.3 of Part I explained that the representation of $\theta \in \mathbb{R}$ is unique except that, for $\theta \in \mathbb{Q} \subset \mathbb{R}$, the final partial quotient a_F , if > 1 , is numerically equal to $(a_F - 1) + \frac{1}{1}$. Therefore one can choose the sequence a_k to end with $a_F - 1, 1$ instead of just a_F . This increases the number of partial quotients by one and thereby inserts an extra ‘quasi-convergent’ in the penultimate place. Since the new final quotient is 1, the recursion relation Eq. 1.2 tells us that the inserted quasi-convergent is

$$\frac{p_k - p_{k-1}}{q_k - q_{k-1}}. \tag{8.6}$$

(You will observe the similarity to Eq 8.5 for three consecutive Farey fraction.) There is therefore the choice of two matrices for any convergent $C_k = p_k/q_k$:

1. \mathcal{M} , which uses the authentic penultimate convergent in the second column, and
2. \mathcal{M}' which uses the inserted quasi-convergent in the second column. Recall that both columns will represent adjacent convergents of some other number with larger denominator (see §2.3).

$$\mathcal{M} = \begin{pmatrix} p_k & p_{k-1} \\ q_k & q_{k-1} \end{pmatrix}, \quad \mathcal{M}' = \begin{pmatrix} p_k & p_k - p_{k-1} \\ q_k & q_k - q_{k-1} \end{pmatrix}. \tag{8.7}$$

I emphasise that in \mathcal{M} , p_{k-1}/q_{k-1} is the authentic penultimate convergent of p_k/q_k . If \mathcal{M} has determinant ± 1 , \mathcal{M}' has determinant ∓ 1 .

Consider the following points based on the example of $\frac{9}{37}$ above:

Example 1 : From the consecutive integers 36, 37 form the matrix $\mathcal{M}_1 = \begin{pmatrix} 9 & 1 \\ 37 & 4 \end{pmatrix}$, with determinant -1 and square $\mathcal{M}_1^2 = \begin{pmatrix} 118 & 13 \\ 481 & 53 \end{pmatrix}$. Note that

1. $\frac{9}{37}$ as $\{0 : 4, 9\}$ has convergents $\frac{1}{4}, \frac{9}{37}$.

2. $\frac{9}{37}$ as $\{0 : 4, 8, 1\}$ has convergents $\frac{1}{4}, \frac{8}{33}, \frac{9}{37}$.

3. $\frac{8}{33} = \{0 : 4, 8\}$.

4. Following item 2, define $\mathcal{M}'_1 = \begin{pmatrix} 9 & 8 \\ 37 & 33 \end{pmatrix}$. As $\mathcal{N} \rightarrow \infty$, the p_k/q_k represented by $\mathcal{M}'_1{}^{\mathcal{N}}\mathcal{J}$ tend to the limit $\theta' = \{0 : 4, \underline{8}, 5\}$, compared with $\theta = \{0 : 4, \underline{13}\}$ above. $8 + 5 = 13$, and the recurring 8 can be identified with $a_F = 8$ in item 3.

Example 2 : From the consecutive integers 51, 52 form the matrix $\mathcal{M}_2 = \begin{pmatrix} 13 & 3 \\ 17 & 4 \end{pmatrix}$, determinant

+1, and square $\mathcal{M}_2^2 = \begin{pmatrix} 220 & 51 \\ 289 & 67 \end{pmatrix}$. Note the following

1. $\frac{13}{17}$ as $\{0 : 1, 3, 4\}$ has convergents $\frac{1}{1}, \frac{3}{4}, \frac{13}{17}$.

2. $\frac{13}{17}$ as $\{0 : 1, 3, 3, 1\}$ has convergents $\frac{1}{1}, \frac{3}{4}, \frac{10}{13}, \frac{13}{17}$.

3. $\frac{10}{13} = \{0 : 1, 3, 3\}$.

4. $\frac{220}{289} = \{0 : 1, 3, \mathbf{5}, \mathbf{3}, 4\}$. Note the insertion of the pair $\{5, 3\}$ before the final partial quotient. This has convergents $\frac{3}{4}, \frac{16}{21}, \frac{51}{67}, \frac{220}{289}$.

5. Some continued fractions associated with higher powers of \mathcal{M}_2 are

$$\mathcal{M}_2^3 : \{0 : 1, 3, \mathbf{5}, \mathbf{3}, \mathbf{5}, \mathbf{3}, 4\}$$

$$\mathcal{M}_2^4 : \{0 : 1, 3, 5, 3, 5, 3, 5, 3, 4\}$$

$$\mathcal{M}_2^8 : \{0 : 1, 3, 5, 3, 5, 3, 5, 3, 5, 3, 5, 3, 5, 3, 5, 3, 4\}$$

Each power inserts a further $\{5, 3\}$ pair. The limiting case is $\{0 : 1, 3, \underline{5}, 3\} \equiv \{0 : 1, \underline{3}, 5\}$.

6. Following Eq 8.7 and item 2, define $\mathcal{M}'_2 = \begin{pmatrix} 13 & 10 \\ 17 & 13 \end{pmatrix}$. As $\mathcal{N} \rightarrow \infty$, $\mathcal{M}'_2{}^{\mathcal{N}}\mathcal{J}$ gives the limit $\theta' = \{0 : 1, 3, \underline{3}, 2, 3\}$. Note how the $\{5, 3\}$ in item 4 has been expanded to $\{3, 2, 3\}$ where $3 + 2 = 5$. The leading '3' can be identified with the final '3' in item 3.

To understand this, first evaluate the limiting value of $\theta = \{0 : 1, \underline{3}, 5\}$ in Example 2 (see item 5 above) by the methods of §3.2.2, Part I:

$$\theta = \frac{1}{1 + \frac{1}{\eta}} \quad \text{where} \quad \eta = 3 + \frac{1}{5 + \frac{1}{\eta}} \quad \text{giving} \quad \theta = \frac{1}{34}(9 + \sqrt{285}) = 0.761237.$$

Next obtained the same result from the matrix form. In the general case define $\mathcal{M} = \begin{pmatrix} p & r \\ q & s \end{pmatrix}$ where $p/q, r/s$ are adjacent convergents and $r = (ps \mp 1)/q$ according to whether the determinant is ± 1 . Suppose that $\mathcal{M}^{\mathcal{N}} = \begin{pmatrix} P & R \\ Q & S \end{pmatrix}$, where \mathcal{N} is large, causing P, Q, R, S all to be very large. The determinant of $\mathcal{M}^{\mathcal{N}}$, however, remains at ± 1 , so $PS \rightarrow QR$ or $P/Q \approx R/S = \theta$, the limiting ratio. Raising \mathcal{M} to the next power will not change this ratio:

$$\begin{pmatrix} p & r \\ q & s \end{pmatrix} \begin{pmatrix} P & R \\ Q & S \end{pmatrix} = \begin{pmatrix} pP + rQ & pR + rS \\ qP + sQ & qR + sS \end{pmatrix} \quad \text{so} \quad \frac{P}{Q} \rightarrow \frac{pP + rQ}{qP + sQ} \rightarrow \theta.$$

Solving for θ

$$\theta = \frac{pq\theta + ps \mp 1}{q^2\theta + qs} \quad \text{from which} \quad \theta = \frac{1}{2q}[(p-s) + \sqrt{(p+s)^2 \mp 4}]. \quad (8.8)$$

As a check, substituting $p = 13$, $q = 17$, $s = 4$ recovers the value of θ in Example 2 above.

Using Eq 8.7, the limiting θ' associated with the alternative matrix $\mathcal{M}' = \begin{pmatrix} p & p-r \\ q & q-s \end{pmatrix}$ is

$$\theta' = \frac{1}{2q}[(p-q+s) + \sqrt{(p+q-s)^2 \pm 4}]. \quad (8.9)$$

Note the change of sign in the root due to the increase in F by 1. As a check, $p = 13$, $q = 17$, $s = 4$ give $\theta' = \sqrt{170}/17 = 0.76696 = \{0 : 1, 3, 3, 2, 3\}$, in agreement with Example 2.

In general each multiplication by \mathcal{M} inserts into the continued fraction a sequence of L partial quotients given by the recursion sequence of θ in Eq 8.8. This in turn inserts L convergents into the continued fraction of $\mathcal{M}^N \mathcal{J}$. This is not unexpected; multiplication by \mathcal{M} must take the continued fraction through one full cycle of recursion, which means insertion of the full sequence, length L , into the a_k .

The only aspect not yet explained is the apparent splitting of one partial quotient on moving from \mathcal{M} to \mathcal{M}' , shown in Examples 1 and 2. This is entirely determined by the continued fraction expansions of Eq 8.8 and 8.9. Some numerical experiments lead me to believe that, where a partial quotient generated by powers of the \mathcal{M} splits under the alternative matrix \mathcal{M}' , then

1. the value of the partial quotient under \mathcal{M} which splits is $\left\lfloor \frac{q}{p} + \left\lfloor \frac{q}{s} \right\rfloor \right\rfloor$,
2. the two partial quotients under \mathcal{M}' into which it splits are $\left\lfloor \frac{q}{s} \right\rfloor - 1$ and $\left\lceil \frac{q}{p} \right\rceil$ respectively, where $\lfloor \cdot \rfloor$ is the floor (lower) integer and $\lceil \cdot \rceil$ the ceiling (higher) integer.

Thus, in Example 2 \mathcal{M}_2 has a partial quotient 5 which splits to 3, 2 under \mathcal{M}'_2 . Here $p = 13$, $q = 17$, $s = 4$, $\left\lfloor \frac{17}{13} + \left\lfloor \frac{17}{4} \right\rfloor \right\rfloor = \left\lfloor \frac{17}{13} + 4 \right\rfloor = 5$. It splits to $\left\lfloor \frac{17}{4} \right\rfloor - 1 = 4 - 1 = 3$ and $\left\lceil \frac{17}{13} \right\rceil = 2$.

Note that in Eq 8.8 $p+s$ is the trace of \mathcal{M} . The expression $\sqrt{Tr^2 - 4}$ is a discriminant in identifying the types of Möbius transformation, which are categorised as hyperbolic, parabolic or elliptic according to whether the trace $Tr > 4$, $= 4$ or < 4 .

Returning to the geometrical interpretation, multiplying \mathcal{J} by powers of a fixed matrix \mathcal{M} sequentially moves the point $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ to discrete points along a trajectory, part of the orbit of $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$. The limit point of this trajectory is the quadratic surd given by Eq 8.7. Indeed, we could regard a quadratic surd as defined as the limit point of $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ under transformation by a matrix formed from one of its convergents.

9 Analysis of the ‘best fit’ property

This section gives the proof of two important theorems about the ‘best fit’ property of the convergents C_k of any real θ in approximating θ . The topic was introduced in Part I §1.6. Write $\varepsilon = |u/v - \theta|$ as the error in u/v . Then the best fit or best approximation property was expressed by saying that to find a fraction u/v such that

$$\varepsilon < \varepsilon_k, \text{ which is equivalent to } \left| \frac{u}{v} - \theta \right| < \left| \frac{p_k}{q_k} - \theta \right|, \text{ we need } v > q_k. \quad (9.1)$$

This is the first theorem to be proved in subsection 9.1. Subsection 9.2 gives a proof of the companion theorem that if a rational number u/v satisfies

$$\varepsilon = \left| \frac{u}{v} - \theta \right| < \frac{1}{2v^2}, \quad \text{Copy of (1.11)}$$

then u/v must be a convergent of θ . This is a very useful test because

- a) it depends only on u/v without the need to find integer or fractional parts, or use relations involving other convergents,
- b) consequently it can be dealt with algebraically, rather than by numerical evaluation.

9.1 Proof that better fit than a convergent implies larger denominator

You may be struck by how similar the best fit property of convergents is to the ‘no intermediate u/v ’ property of adjacent fractions a/b , c/d in a Farey series, for which $ad - bc = \pm 1$, described in §8.2. Since θ lies between any two of its adjacent convergents, and since there is no intervening fraction u/v with intermediate denominator $q_{k-1} < v < q_k$, any u/v with intermediate denominator much lie further away from θ than either C_{k-1} or C_k , and so have larger error. Personally, I find this a satisfactory and sufficient explanation. However, a more thorough and probably more elegant treatment is given by Hardy and Wright, so here is my expanded account of their argument.

The theorem in Eq 9.1 addresses the denominators of the convergents, and so any proof must also deal with denominators. Hardy and Wright’s proof proceeds in three stages. They establish that for $q_{k-1} < v < q_k$

1. $|p_k - \theta q_k| < |p_{k-1} - \theta q_{k-1}|$, which is $\varepsilon_k q_k < \varepsilon_{k-1} q_{k-1}$,
2. $|p_k - \theta q_k| < |u - \theta v|$, which is $\varepsilon_k q_k < \varepsilon v$,
3. $\left| \frac{p_k}{q_k} - \theta \right| < \left| \frac{u}{v} - \theta \right|$ which is $\varepsilon_k < \varepsilon$.

Item 2 is a stronger statement than item 3, so 3 is just a corollary of item 2.

Step 1 : From Eqs 1.3 and 1.5

$$\theta - \frac{p_k}{q_k} = \frac{\theta_{k+1} p_k + p_{k-1}}{\theta_{k+1} q_k + q_{k-1}} - \frac{p_k}{q_k} = \frac{(-1)^k}{q_k (\theta_{k+1} q_k + q_{k-1})}$$

$$\text{so } |\theta q_k - p_k| \equiv \varepsilon_k q_k = \frac{1}{\theta_{k+1} q_k + q_{k-1}}. \quad (9.2a)$$

Reducing the index by 1

$$|\theta q_{k-1} - p_{k-1}| \equiv \epsilon_{k-1} q_{k-1} = \frac{1}{\theta_k q_{k-1} + q_{k-2}}. \quad (9.2b)$$

Now use the recursion relation to substitute for q_k in Eq 9.2a. Also put $\theta_k = a_k + 1/\theta_{k+1}$ (from Eq 1.1) into Eq 9.2b:

$$\begin{aligned} \epsilon_k q_k &= \frac{1}{(\theta_{k+1} a_k + 1) q_{k-1} + \theta_{k+1} q_{k-2}} \\ \epsilon_{k-1} q_{k-1} &= \frac{1}{(a_k + \frac{1}{\theta_{k+1}}) q_{k-1} + q_{k-2}} \end{aligned}$$

Now $\theta_{k+1} > 1$ so $\theta_{k+1} q_{k-2} > q_{k-2}$. Also $\theta_{k+1} a_k + 1 > a_k + 1 > a_k + \frac{1}{\theta_{k+1}}$. Thus the denominator of $\epsilon_k q_k$ is greater than the denominator of $\epsilon_{k-1} q_{k-1}$. We have established that $\epsilon_k q_k < \epsilon_{k-1} q_{k-1}$. As a corollary

$$\epsilon_k < \epsilon_{k-1} \frac{q_{k-1}}{q_k} < \epsilon_{k-1} \quad \text{with} \quad \frac{q_{k-1}}{q_k} \approx \frac{1}{a_k + \frac{1}{a_{k-1}}}.$$

Step 2 : It is possible to express u/v as a linear rational combination of the two convergents whose denominators straddle v :

$$\frac{u}{v} = \frac{\alpha p_k + \beta p_{k-1}}{\alpha q_k + \beta q_{k-1}}, \quad \alpha, \beta \in \mathbb{Z}, \quad \alpha, \beta \neq 0, \quad \text{gcd}(\alpha, \beta) = 1.$$

It is correct to identify u with the numerator and v with the denominator. α and β can be found by matrix inversion:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} p_k & p_{k-1} \\ q_k & q_{k-1} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad \text{so} \quad \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} q_{k-1} & -p_{k-1} \\ -q_k & p_k \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}$$

or

$$\alpha = u q_{k-1} - v p_{k-1}, \quad \beta = -u q_k + v p_k.$$

Now here is the key step : since v lies between q_{k-1} and q_k , the integer coefficients α and β cannot both be positive – they must have opposite signs. But $p_{k-1} - \theta q_{k-1}$ and $p_k - \theta q_k$ also have opposite signs because adjacent convergents lie either side of θ . Therefore the products

$$\alpha(p_k - \theta q_k), \quad \beta(p_{k-1} - \theta q_{k-1})$$

have the same sign. When these are added

$$\alpha(p_k - \theta q_k) + \beta(p_{k-1} - \theta q_{k-1}) = (\alpha p_k + \beta p_{k-1}) - \theta(\alpha q_k + \beta q_{k-1}) = u - \theta v.$$

Since $|\alpha| \geq 1$, $|\beta| \geq 1$, and since the terms in this sum reinforce whatever their sign,

$$|p_k - \theta q_k| < |p_{k-1} - \theta q_{k-1}| < |u - \theta v|. \quad (9.3a)$$

We have established that

$$\epsilon_k q_k < \epsilon_{k-1} q_{k-1} < \varepsilon v. \quad (9.3b)$$

Step 3 : Divide Eq 9.3a or b by q_k and use $v < q_k$:

$$\epsilon_k < \varepsilon \frac{v}{q_k} \text{ so in all cases } \epsilon_k < \varepsilon,$$

which proves the complement of Eq 9.1, and so the inequality Eq 9.1 itself.

Step 3, however, leaves it unclear how ϵ_{k-1} relates to ϵ . Eq 9.3 merely states that $\epsilon_{k-1}q_{k-1}/v < \epsilon$, but $\epsilon_{k-1}q_{k-1}/v < \epsilon_{k-1}$ also. However, the ‘no intermediate u/v ’ property of Farey series argues strongly that $\epsilon_{k-1} < \epsilon$, and this is borne out by numerical evidence as in the example below.

Numerical example Take $\theta = 0.3927$. Two adjacent convergents are

$$p_{k-1}/q_{k-1} = 9/23 = 0.3913 \text{ (odd) with error } -14 \times 10^{-4}, \text{ and}$$

$$p_k/q_k = 11/28 = 0.3929 \text{ (even) with error } +1.6 \times 10^{-4}.$$

Suppose we take $v = 26$ since $23 < 26 < 28$. We know from the Farey series discussion in §8.2 that there is no fraction $u/26$ between these two convergents, and indeed the closest such fraction to θ is $u/v = 10/26 = 5/13 = 0.3846$. For this $\alpha = -2$, $\beta = 3$, illustrating the opposite signs. The error in $5/13$ is -81×10^{-4} , much bigger than that in even the lower convergent.

Moreover, the errors agree with the formula : $(-2 \times 1.6 \times 28 + 4 \times (-14) \times 23)/13 = -81$, all $\times 10^{-4}$.

Let’s take a significantly larger denominator : $u/v = 19/48 = 0.3958$. For this $\alpha = 5$, $\beta = -4$, the signs being opposite again. The error is -31×10^{-4} , 20 times poorer than the convergent $11/28$ which has a lower denominator. So this is another instance of ϵ exceeding the error at a convergent with lower denominator.

9.2 Proof of $< 1/2v^2$ test for a convergent

Part I, and particularly §1.6, described the ‘best fit’ property of the convergents C_k of any real θ in approximating θ . A formal statement of the theorem (Hardy and Wright, page 140) is

Suppose that θ and ξ are related by

$$\theta = \frac{\xi A + C}{\xi B + D} \qquad \text{Copy of (8.1)}$$

where $\xi > 1$ (strictly), A, B, C, D are positive integers, $AD - BC = \pm 1$, and $B > D$. Then the fraction A/B is the k^{th} convergent and C/D the $(k-1)^{\text{th}}$ convergent of θ , and ξ is the $(k+1)^{\text{th}}$ complete quotient, and the function corresponds to Eq 1.3.

Key to the proof is to have a sufficient list of the essential, defining properties of convergents to test whether any given fraction is in fact a convergent. For this purpose Eqs 1.1 and 1.3 are taken as the defining criteria and from them the proof is in two stages:

1. to list the defining properties of convergents,
2. to show that u/v with $\epsilon < 1/2v^2$ possesses all these defining properties.

9.2.1 Defining features of a convergent

Let’s place the defining properties of convergents within the context on mappings and matrices in §8. Convergents are equivalent numbers subject to the further restrictions that in Eq 8.1 the integer constants in the matrix $\begin{pmatrix} A & C \\ B & D \end{pmatrix}$ satisfy $B > D$. This latter restriction implies that $A > C$ ⁹.

The steps in the proof are as follows:

⁹Let $B = D + \delta$. Then $D(A - C) = C\delta \pm 1 > 0$ except if $C = 1, \delta = 1$, which case we can ignore. Hence $A > C$.

1. Since A/B is a common fraction, it can be expanded into a finite continued fraction $\{a_0 : a_1, a_2, \dots, a_F\}$ or $\{a_0 : a_1, a_2, \dots, a_F - 1, 1\}$. The final convergent of this is $C_F = p_F/q_F = A/B$, and the penultimate is p_{F-1}/q_{F-1} . Now $\gcd(p_F, q_F) = 1$, this being a property of all convergents. In addition $\gcd(A, B) = 1$ because, if it were some integer $j > 1$ such that $B = jA$, the determinant would be $AD - jAC$; however this can be ± 1 only if $A = 1$, which is not generally the case. We therefore conclude that $A = p_F$ and $B = q_F$.
2. From Eq 1.5 $p_F q_{F-1} - p_{F-1} q_F = (-1)^{F-1}$. Equate this to the determinant, $AD - BC$. To match the + and - signs does require that F be appropriately odd or even, depending on $AD - BC$. However, this can be achieved by choosing the appropriate continued fraction, ending either in a_F or $(a_F - 1), 1$. We thus have

$$p_F q_{F-1} - p_{F-1} q_F = AD - BC = p_F D - q_F C \quad \text{so} \quad p_F (q_{F-1} - D) = q_F (p_{F-1} - C) .$$

3. To satisfy this last equation

- either the left side is a non-zero multiple of q_F and the right side a multiple of p_F ,
- or both $(q_{F-1} - D)$ and $(p_{F-1} - C)$ are zero, in which case $q_{F-1} = D$ and $p_{F-1} = C$.

The first bullet point would require $q_F \mid (q_{F-1} - D)$, since $\gcd(p_F, q_F) = 1$. But this is not possible because $|q_{F-1} - D| < q_F$. This is because q_{F-1} and D are both positive integers, both less than q_F ; clearly $q_{F-1} < q_F$ because it is from a lower convergent, and $D < B = q_F$ is a condition we imposed. Similarly $p_F \nmid (p_{F-1} - C)$. Hence only the second bullet point is possible. We have established that C/D is the penultimate convergent of A/B .

4. The identification $\xi = \theta_{k+1} > 1$ is made by comparing Eq 9.1 with Eq 1.3. As a continued fraction either $\theta = \{a_0 : a_1, a_2, \dots, a_F, \xi\}$ or $\theta = \{a_0 : a_1, a_2, \dots, a_F - 1, 1, \xi\}$. A choice of any real $\xi \geq 1$ then specifies θ (there are infinite possibilities), but $A/B, C/D$ will certainly be adjacent convergents of that θ .
5. A comment is needed on the condition $\xi > 1$, strictly. If $\xi = 1$, it must be the last partial quotient a_F of a rational number. Following the discussion in §2.3, Part I, as such it should be added to the penultimate partial quotient. Doing so removes the ‘extra’ quasi-convergent generated by the extra partial quotient. The effect of leaving $\xi = 1$ is that the fraction C/D could be either the quasi-convergent, listed in the position of C_{F-1} , or the authentic penultimate convergent, appearing in the position of C_{F-2} . This is the only ambiguous situation.

Numerical example From the consecutive integers $51 = 3 \times 17$ and $52 = 4 \times 13$, set $A/B = 13/17$, $C/D = 3/4$, which satisfy $B > D, A > C$. $13/17 = \{0 : 1, 3, 4\}$, with three convergents $1/1, 3/4, 13/17$. $F = 3$ is odd so $(-1)^{F-1} = +1$. Now introduce an irrational as ξ . I choose $\xi = \pi/2 \approx 1.570796 = \{1 : 1, 1, 3, 31, 1, 145, \dots\}$, so

$$\theta = \frac{1 \cdot 570796 \times 13 + 3}{1 \cdot 570796 \times 17 + 4} = 0.762790\dots$$

As a continued fraction this is $\{0 : 1, 3, 4, 1, 1, 1, 3, 31, 1, 145, \dots\}$ and has convergents $1/1, 3/4, 13/17, 16/21, 29/38$, etc. Our two chosen fractions $A/B, C/D$ feature here, and in the same positions as in the expansion of A/B . So no surprises here!

9.2.2 Best fit of u/v to θ implies a convergent

Now for the second part of the proof. We are given $\theta \in \mathbb{R}$ and u/v in lowest terms such that

$$\left| \frac{u}{v} - \theta \right| < \frac{1}{2v^2}. \quad \text{Copy of (1.11)}$$

We aim to prove that u/v is a convergent of θ .

This proof is the reverse of the proof of the theorem in §9.1. It starts by expanding u/v as a finite continued fraction, and evaluating the final convergent $p_F/q_F = u/v$ and the penultimate one p_{F-1}/q_{F-1} . Being adjacent convergents, they satisfy Eq 1.5: *viz:* $|uq_{F-1} - p_{F-1}v| = 1$. The associated matrix has determinant ± 1 and so is invertible, as in a Möbius transformation. Hence there exists a ξ equivalent to θ (§8.3) such that

$$\theta = \frac{\xi p_F + p_{F-1}}{\xi q_F + q_{F-1}}.$$

Whether p_F/q_F and p_{F-1}/q_{F-1} are convergents of θ (as well as of u/v) will depend on whether all criteria listed in §9.2.1 are met. Since we have the constants being integers and determinant being ± 1 , the outstanding criterion is that, strictly, $\xi > 1$. To determine this write Eq 1.11 in terms of the convergents of u/v :

$$\begin{aligned} \left| \frac{u}{v} - \frac{\xi u + p_{F-1}}{\xi v + q_{F-1}} \right| &< \frac{1}{2v^2}, \\ \frac{|\xi uv + uq_{F-1} - \xi vu - vp_{F-1}|}{\xi v + q_{F-1}} &< \frac{1}{2v}, \\ \frac{1}{\xi v + q_{F-1}} &< \frac{1}{2v} \\ \xi v + q_{F-1} &> 2v \\ \xi &> 2 - \frac{q_{F-1}}{v} > 1 \end{aligned}$$

because $v = q_F$ and $q_{F-1} < q_F$. Thus all criteria for u/v to be a convergent of θ are met.

9.3 Are the convergents of θ its only good approximations?

It is well established from the analysis in §1.7, Part I, and in §9.2 above that the convergent $C_k = p_k/q_k$ of any real θ is the best rational approximation to θ of all u/v for $v \leq q_k$. But suppose we set a higher limit on denominator, v_{max} such that $q_k < v_{max} < q_{k+1}$; could there be a better approximation than C_k ? I believe the answer to be Yes in certain circumstances.

To illustrate this I have carried out a small numerical investigation using $\theta = \{1 : 2, 3, 4, 5, 6\}$. I listed all fractions u/v which are at all close to θ in order of increasing denominator v , and calculated the difference $u/v - \theta$, noting those u/v at which the error decreased to the lowest value so far. I thereby find that there are several u/v which are not convergents of θ , but which do nevertheless give step improvements in accuracy. Table 19 below lists those fractions which give a better approximation to $\{1 : 2, 3, 4, 5, 6\}$ than all fractions with smaller denominator. Each remains the best approximation until the next fraction in the table is reached.

It is pretty clear what is happening. Using the terminology of §2.3, the authentic convergents are those in which the continued fraction for θ has one or more of its least significant a_k deleted.

Table 19: Fractions which approximate best to $\theta = \{1: 2, 3, 4, 5, 6\}$

Fraction	Error, ϵ_k	Sign	Convergent?	Expansion
1/1	-0.4331276	-	Yes : C_0	{ 1 : }
3/2	0.0668724	+	Yes : C_1	{ 1 : 2 }
7/5	-0.0331276	-	No	{ 1 : 2, 2 }
10/7	-0.0045561	-	Yes : C_2	{ 1 : 2, 3 }
23/16	0.0043724	+	No	{ 1 : 2, 3, 2 }
33/23	0.0016550	+	No	{ 1 : 2, 3, 3 }
43/30	0.0002058	+	Yes : C_3	{ 1 : 2, 3, 4 }
139/97	-0.0001379	-	No	{ 1 : 2, 3, 4, 3 }
182/127	-0.0000567	-	No	{ 1 : 2, 3, 4, 4 }
225/157	-0.0000066	-	Yes : C_4	{ 1 : 2, 3, 4, 5 }
718/501	0.0000062	+	No	{ 1 : 2, 3, 4, 5, 3 }
943/658	0.0000031	+	No	{ 1 : 2, 3, 4, 5, 4 }
1168/815	0.0000013	+	No	{ 1 : 2, 3, 4, 5, 5 }
1393/972	0		Yes : C_F	{ 1 : 2, 3, 4, 5, 6 }

The intervening fractions u/v correspond to quasi-convergents in which the final partial quotient, a_F , of the next convergent ($q_k > v$) is replaced by $a_F - 1$, $a_F - 2$, $a_F - 3$, etc. Figure 6 illustrates the positions of two convergents and two quasi-convergents relative to θ on the real number line. The case illustrated is $\theta = \{1: 2, 3, 4, 5, 6\}$, $C_{k-1} = \{1: 2, 3\} = \frac{10}{7}$, $C_k = \{1: 2, 3, 4\} = \frac{43}{30}$ with quasi-convergents $\{1: 2, 3, 3\} = \frac{33}{23}$ and $\{1: 2, 3, 2\} = \frac{23}{16}$. Consistent with Eq 8.5 derived for three consecutive Farey fractions, the quasi-convergent $(p_k - p_{k-1})/(q_k - q_{k-1}) = \frac{33}{23}$ lies to the right of C_k , whilst C_{k-1} lies to the left. Both $\{1: 2, 3, 3\}$ and $\{1: 2, 3, 2\}$ (just!) are closer to θ than C_{k-1} . However $\{1: 2, 3, 1\}$ (not shown) lies even further to the right and has greater error.

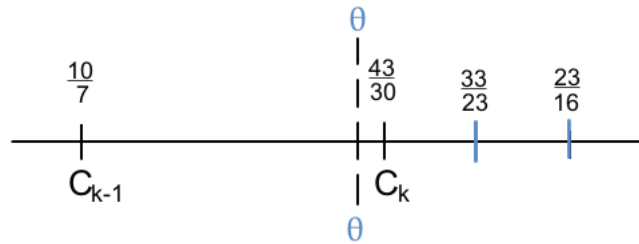


Figure 6: The real number line showing positions of convergents and quasi-convergents

There is a pleasing geometric model which makes the difference between authentic and quasi-convergents very clear. It was originally proposed (for authentic convergents only) by Felix Klein. You plot each as a point on cartesian axes, with the denominator as the x co-ordinate and the numerator as y co-ordinate. Figure 7 is a not-to-scale schematic. All the points are on a lattice of integer co-ordinates. The numerical value of C_k corresponds to the gradient of the line from the origin to the point (q_k, p_k) . C_0 is the integer part (which in this diagram is zero). All even convergents under-estimate θ so lie nearer the x -axis, while all odd ones give over-estimates.

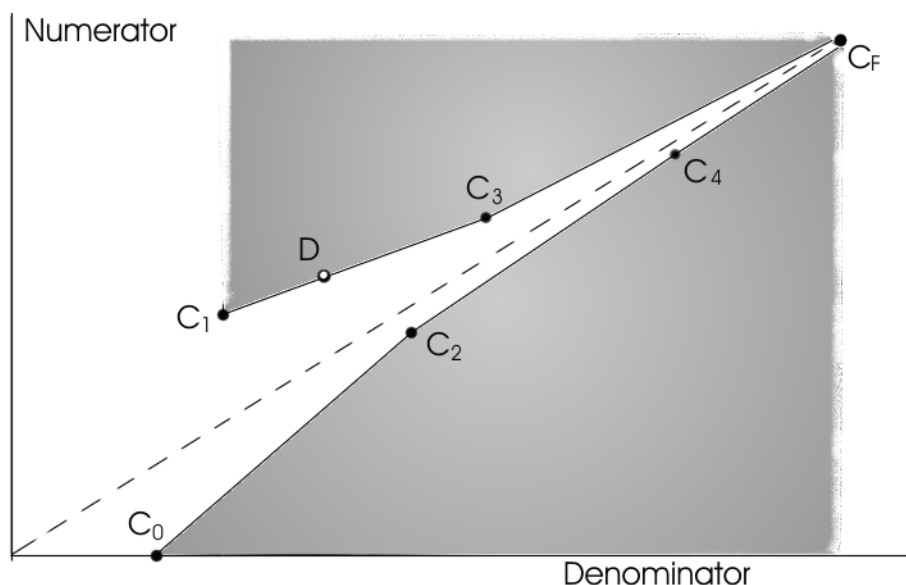


Figure 7: Illustrating the convergents to θ as vertices of left and right polygons

In Table 19 separate values into those with a + error and those with -. Within either the + or - category, choose any fraction as starting point and imagine the straight line segment drawn to the next-but-one fraction: for example, from (1,1) to (7,10). Calculate the y (numerator) value on this line corresponding to the x (denominator) value of the intermediate fraction. At $x = 5$ the line from (1, 1) to (7, 10) passes through (5, 7). This corresponds exactly to the intermediate fraction $7/5$, a quasi-convergent.

Now repeat the procedure where start and end points straddle a convergent; for example, from (5, 7) to (97, 139). On this line at $x = 7$, the y value would be $9 \cdot 8696$. However there is a convergent at denominator 7 with value precisely 10. (7, 10) therefore lies outside the line segment between (5, 7) and (97, 139). This means that the line segments joining the + fractions form an open polygon, the vertices of which lie at the + convergents to θ , and similarly for the - fractions. The convergents, therefore, are at the vertices while points like D, lying on a straight line between adjacent vertices, correspond to quasi-convergents. The positioning of quasi-convergents on such lines between two authentic convergents is consistent with Eq 8.5 for three consecutive fractions in a Farey sequence, where the numerator (or denominator) of the middle fraction is a linear combination of the numerators (or denominators) to either side.

You could make an actual model like Figure 7 by knocking nails into a wooden board at positions corresponding to θ and its convergents, tying a string from C_0 to θ and pulling it tight. A second string can be pulled from C_1 to θ to give the upper and lower boundary lines in Figure 7. Readers familiar with linear programming may also see a similarity with the theorem that any linear function defined on a convex polytope attains its maximum and minimum values precisely at the vertices.

The $1/(2 \times \text{denominator}^2)$ test of §9.2 would seem to be a definitive test to distinguish quasi-convergents from authentic ones. Just to check this, apply the methods of §9.2 to the quasi-convergent

$(p_k - p_{k-1})/(q_k - q_{k-1})$. The error ϵ_Q is

$$\begin{aligned}\epsilon_Q &\equiv \left| \frac{p_k - p_{k-1}}{q_k - q_{k-1}} - \theta \right| = \left| \frac{p_k - p_{k-1}}{q_k - q_{k-1}} - \frac{\theta_{k+1}p_k + p_{k-1}}{\theta_{k+1}q_k + q_{k-1}} \right| \\ &= \frac{\theta_{k+1} + 1}{(\theta_{k+1}q_k + q_{k-1})(q_k - q_{k-1})}\end{aligned}$$

which is to be compared with $1/(q_k - q_{k+1})^2$. The ratio is

$$\frac{(\theta_{k+1} + 1)(q_k - q_{k-1})}{\theta_{k+1}q_k + q_{k-1}}$$

and we want to know whether this can be less than 1 or even $\frac{1}{2}$ – we hope not! To assess this, write q_k as a multiple of q_{k-1} , namely $q_k = Aq_{k-1}$, $A > 1$. We also know that $\theta_{k+1} > 1$. The ratio becomes

$$\frac{(\theta_{k+1} + 1)(A - 1)}{A\theta_{k+1} + 1}.$$

This is less than 1 for $A < \approx 5$, which will be the majority of cases, showing how well quasi-convergents can mimic authentic ones. The ratio *can* be less than $\frac{1}{2}$. At the limit $\theta_{k+1} = 1$ this happens if $A < \frac{5}{3}$, whilst in the other limit $\theta_{k+1} \rightarrow \infty$ it happens if $A < 2$. Of course, to obtain $A < 2$ means that $a_k = 1$, $q_k = q_{k-1} + q_{k-2}$. But if the last convergent $a_F = 1$, the convergent p_k/q_k is authentic. Quasi-convergents only arise if $a_F \geq 2$, so that it can be split into $\{a_F - 1, 1\}$. Therefore a quasi-convergent cannot ‘squeeze through the net’ of the $1/(2 \times \text{denominator}^2)$ test.

We are now in a position to answer the question

Challenge For a given maximum allowed value of denominator v_{max} , what is the best rational approximation to a given real θ ?

The answer is trivial if $\theta = P/Q \in \mathbb{Q}$ and $Q < v_{max}$: it is just P/Q itself. For θ irrational, or rational with $v_{max} < Q$, the answer is found as follows:

Recipe Determine the continued fraction expansion for θ as far as the first convergent for which $q_k > v_{max}$. Let the final partial quotient of this C_k be a_F . Replace a_F in turn by $a_F - 1$, $a_F - 2$, $a_F - 3$, etc. and evaluate each as an ordinary fraction. Check the error of each and select the one with i) highest denominator $< v_{max}$ and ii) a smaller error than C_{k-1} – (IF one exists! No guarantee!) .

Numerical example Find the best rational approximation u/v to $\ln 3 \approx 1.09861229$ which has a denominator $v \leq 400$.

Method : Using a sufficiently precise decimal value for $\ln 3$, determine the continued fraction to be $\{1: 10, 7, 9, 2, 2, 1, \dots\}$. The first few convergents up to $q_k > v_{max}$ are

$$1, \quad \frac{11}{10}, \quad \frac{78}{71}, \quad \frac{713}{649}, \quad \text{etc.}$$

The next $q_k > v_{max}$ is 649, corresponding to $\{1: 10, 7, 9\}$, for which $a_F = 9$. So replace this by 8, 7, 6, etc. and evaluate until $v < v_{max}$. Test the error against the error at the lower convergent, which is $78/71 - \ln 3 = -2.07 \times 10^{-5}$.

$$1. \{1: 10, 7, 8\} = 635/578, \text{ with } 578 > v_{max} = 400 : 635/578 - \ln 3 = 0.4 \times 10^{-5}$$

$$2. \{1: 10, 7, 7\} = 557/507, \text{ with } 507 > v_{max} : 557/507 - \ln 3 = 0.7 \times 10^{-5}$$

3. $\{1: 10, 7, 6\} = 479/436$, with $436 > v_{max}$: $479/436 - \ln 3 = 1.16 \times 10^{-5}$

4. $\{1: 10, 7, 5\} = 401/365$, with $365 < v_{max}$: $401/365 - \ln 3 = 1.78 \times 10^{-5}$

This last fraction meets the two criteria of i) $v < 400$, and ii) error less (marginally!) than at the lower convergent $78/71$. It will be clear, however, that the errors are increasing, and indeed at $\{1: 10, 7, 4\} = 323/294$ it has increased to 2.72×10^{-5} , larger than at $78/71$. This shows again just how sharp and efficient the convergents are as approximations. One cannot be sure of finding a more accurate quasi-convergent with $v < v_{max}$.

10 Convergence of sequences C_k , ϵ_k for quadratic irrationals

This section is about the rate at which the sequence of convergents C_k converges to its limit θ . This is equivalent to quantifying the rate at which the error $\epsilon_k \rightarrow 0$ as $k \rightarrow \infty$. It therefore takes further some topics introduced in §1.7 and partially analysed in §9. The central result is that the sequence converges exponentially, as a geometric series, with modest variation from this depending on the particular values of the partial quotients, a_k .

It will be obvious that the larger the values of the partial quotients a_k , the more quickly the series C_k will converge to its final value (finite fraction) or limiting value (infinite fraction), θ . This section examines the trend for quadratic irrationals. Higher order irrationals are considered in §11 and 12. For quadratics the behaviour depends on the recurrence lengths L of the a_k . For $L = 1$ and 2 the convergents tend asymptotically to a single geometric series as $k \rightarrow \infty$. If L is 3 or longer, the trend in convergence is as a geometric series but with a small superimposed periodic fluctuation, created by the interleaving of three or more (respectively) geometric series with the same common ratio but different initial values. For recurring continued fractions it is possible not only to quantify the asymptotic geometric series, but to find exact closed form expressions for the numerators p_k and denominators q_k , and the error ϵ_k .

10.1 Asymptotic convergence of $\{a : \underline{a}\}$, $L = 1$

10.1.1 Trend to a geometric series

The numerical example in Table 20 shows the behaviour we will be investigating in this section. It lists the first 18 convergents of $\{2 : \underline{2}\}$, their differences Δ_k as logarithms, and the differences in these logarithms as k increases. They become stable, at -1.762747174 , which is $\ln(|-3 + 2\sqrt{2}|)$. This is the behaviour of a geometric series with common ratio $-3 + 2\sqrt{2}$.

To analyse this consider first the behaviour as $k \rightarrow \infty$ of the simplest case $\theta = \{a : \underline{a}\} > 0$. From §3.2.1 we know that

$$\theta^2 - a\theta - 1 = 0 \quad \text{with solutions} \quad \theta = \frac{1}{2}(a + \sqrt{a^2 + 4}), \quad \frac{-1}{\theta} = \frac{1}{2}(a - \sqrt{a^2 + 4}). \quad (10.1)$$

To simplify notation, introduce σ by

$$\sigma = \sqrt{a^2 + 4}, \quad \text{making} \quad \theta = \frac{1}{2}(a + \sigma), \quad \frac{-1}{\theta} = \frac{1}{2}(a - \sigma).$$

In §1.5, Eq 1.6 Δ_k was defined as the difference of adjacent convergents, $C_k - C_{k-1}$ for $k \geq 1$. We focus on the ratio

$$\frac{\Delta_k}{\Delta_{k-1}} = \frac{-q_{k-2}}{q_k}.$$

As $k \rightarrow \infty$

$$\frac{-q_{k-2}}{q_k} = \frac{-q_{k-2}}{a q_{k-1} + q_{k-2}} = \frac{-1}{a \left(\frac{q_{k-1}}{q_{k-2}} \right) + 1} = \frac{-1}{a \left(\frac{p_{k-2}}{q_{k-2}} \right) + 1} \rightarrow \frac{-1}{a\theta + 1} = R, \quad (10.2)$$

the limiting common ratio. Substituting for θ

$$R \equiv \lim_{k \rightarrow \infty} \frac{\Delta_k}{\Delta_{k-1}} = \frac{-\sigma - a}{\sigma + a} = -\frac{1}{2}(a^2 + 2 - a\sigma). \quad (10.3)$$

Note that $|R| < 1$ and is independent of k . This means that the difference Δ_k between convergents tends to zero as a geometric series

$$\Delta_k \approx \tilde{\Delta}_1 R^{k-1}, \quad k \rightarrow \infty \quad (10.4)$$

Table 20: Logarithm of differences Δ_k between successive convergents of $\{2 : \underline{2}\}$

k	C_k	$-\ln \Delta_k $	$\frac{\ln \Delta_k - \ln \Delta_{k-1} }{\ln \Delta_k / \Delta_{k-1} } =$
0	2 / 1		
1	5 / 2	-0.693147181	
2	12 / 5	-2.302585093	-1.609437912
3	29 / 12	-4.094344562	-1.791759469
4	70 / 29	-5.852202480	-1.757857918
5	169 / 70	-7.615791072	-1.763588592
6	408 / 169	-9.378393957	-1.762602885
7	985 / 408	-11.14116589	-1.762771932
8	2378 / 985	-12.90390882	-1.762742926
9	5741 / 2378	-14.66665672	-1.762747902
10	13860 / 5741	-16.42940377	-1.762747048
11	33461 / 13860	-18.19215098	-1.762747217
12	80782 / 33461	-19.95489819	-1.762747206
13	195025 / 80782	-21.71764526	-1.762747071
14	470832 / 195025	-23.48039593	-1.762747174
15	1136689 / 470832	-25.24313966	-1.762747174
16	2744210 / 1136689	-27.00588683	-1.762747174
17	6625109 / 2744210	-28.76863400	-1.762747174
18	15994428 / 6625109	-30.53138118	-1.762747174

where $\tilde{\Delta}_1$ is a notional first term in the asymptotic series. The tilde character $\tilde{}$ indicates that this is not an actual difference, but only an extrapolation of the asymptotic series back to $k = 0$. Similarly, introduce notional convergents \tilde{C}_k related to the geometric series. These correspond to the $\tilde{\Delta}_k$ through $\tilde{C}_k - \tilde{C}_{k-1} = \tilde{\Delta}_k$ and tend to the true convergents as k increases. In terms of these

$$\Sigma_\infty = \lim_{F \rightarrow \infty} (\tilde{C}_1 - \tilde{C}_0) + (\tilde{C}_1 - \tilde{C}_2) + \dots + (\tilde{C}_F - \tilde{C}_{F-1}) = -\tilde{C}_0 + \theta$$

where subscript F denotes a final convergent, which tends to θ . \tilde{C}_0 is the notional first term in the sequence of approximate convergents, yet to be determined. The sum to infinity of this series is

$$\Sigma_\infty = \tilde{\Delta}_1 + \tilde{\Delta}_2 + \dots + \tilde{\Delta}_k + \dots = \frac{\tilde{\Delta}_1}{1 - R}.$$

Essentially we now have an approximation to any convergent C_k , and to any difference of convergents Δ_k , in terms of an asymptotic geometric series with known ratio R and unknown parameters \tilde{C}_0 and $\tilde{\Delta}_1$. These latter are related through $(1 - R)(\theta - \tilde{C}_0) = \tilde{\Delta}_1$. Their values will be determined shortly, after looking at the convergence of $\{a : \underline{a}\}$ from another point of view.

10.1.2 Linear difference equation

Having set the scene, let us now analyse $\{a : \underline{a}\}$ by treating the recursion relations for q_k and p_k , Eq 1.2, as linear difference equations with constant coefficients, 1 and a respectively:

$$q_{k+2} = aq_{k+1} + q_k.$$

There is a well established method for solving such equations. To remind the reader, let \mathcal{D} be the operator which transforms q_k to q_{k+1} *i.e.* $\mathcal{D}(q_k) = q_{k+1}$. Then the recursion relation can be written

$$\mathcal{D}(\mathcal{D}(q_k)) - a\mathcal{D}(q_k) - q_k = 0 \quad \text{or} \quad (\mathcal{D}^2 - a\mathcal{D} - 1)q_k = 0.$$

Associated with this operator equation is the characteristic or auxiliary equation $\lambda^2 - a\lambda - 1 = 0$. But this is has exactly the same form as Eq 10.1, used above to evaluate θ , so immediately we can identify λ with θ . The roots of the auxiliary equation are θ and $-1/\theta$. Treating \mathcal{D} in a purely formal manner, this quadratic can be factorised as

$$(\mathcal{D} - \theta) \left(\mathcal{D} + \frac{1}{\theta} \right) q_k = 0$$

The equation is satisfied if either linear factor equals 0, that is if either

$$(\mathcal{D} - \theta)q_k = 0 \quad \text{or} \quad \left(\mathcal{D} + \frac{1}{\theta} \right) q_k = 0, \quad \text{meaning that}$$

$$q_{k+1} = \theta q_k \quad \text{or} \quad q_{k+1} = -\frac{1}{\theta} q_k.$$

Each of these describes a geometric series with common ratios θ and $-1/\theta$ respectively. If both series started from some initial value, which can be taken as 1, they would give the k^{th} term to be

$$q_k = \theta^k \quad \text{and} \quad q_k = \frac{(-1)^k}{\theta^k}.$$

The general solution of the difference equation must be an arbitrary linear sum of these two independent solutions:

$$q_k = K_1 \theta^k + K_2 \frac{(-1)^k}{\theta^k}.$$

Now choose the multipliers K_1, K_2 to fit the first two convergents, $q_0 = 1, q_1 = a$:

$$\text{Solution for } q_k: \quad K_1 = \frac{\sigma + a}{2\sigma} = \frac{\theta}{\sigma} \quad \text{and} \quad K_2 = \frac{\sigma - a}{2\sigma} = -\frac{1}{\sigma\theta}. \quad (10.5a)$$

There are several ways these can be written; an alternative is

$$K_1 = \frac{a\theta + 1}{\theta^2 + 1}, \quad K_2 = \frac{\theta(\theta - a)}{\theta^2 + 1}. \quad (10.5b)$$

Let $M_{1,2}$ be the corresponding coefficients for p_k . These are obtained by fitting to $p_0 = a, p_1 = a^2 + 1$:

$$M_1 + M_2 = a \quad \text{and} \quad M_1 \theta - \frac{M_2}{\theta} = a^2 + 1$$

$$\text{with solution for } p_k: \quad M_1 = \frac{a^2 + 2 + a\sigma}{2\sigma} \quad \text{and} \quad M_2 = \frac{-a^2 - 2 + a\sigma}{2\sigma}. \quad (10.5c)$$

$$\text{or alternatively} \quad M_1 = \theta - \frac{1}{\sigma} = \frac{\theta^2}{\sigma} \quad \text{and} \quad M_2 = -\frac{1}{\theta} + \frac{1}{\sigma} = -\frac{1}{\sigma\theta^2} \quad (10.5d)$$

Equations 10.5 are known as the Binet form of solution of the difference equation, after the nineteenth century French mathematician Jacques Binet.

You can check that in Eqs 10.5 $k = 0$ gives $p_0 = a, q_0 = 1$, followed by $p_1 = a^2 + 1, p_2 = a^3 + 2a, p_3 = a^4 + 3a^2 + 1$, etc. and $q_1 = a, q_2 = a^2 + 1$, etc. in complete agreement with applying the recursion

relations directly. Since in this case $q_k = p_{k-1}$, explicit expressions for the p_k are also obtained merely by increasing the power k by 1 in the formula for q_k ; that is, $M_1 = K_1\theta$, $M_2 = K_2/\theta$. Consequently the product $K_1K_2 = M_1M_2 = -1/\sigma^2$.

The exact expression for convergent C_k is

$$C_k = \frac{p_k}{q_k} = \frac{K_1\theta^{k+1} + K_2\left(\frac{-1}{\theta}\right)^{k+1}}{K_1\theta^k + K_2\left(\frac{-1}{\theta}\right)^k} = \frac{\theta + \frac{K_2}{K_1}\frac{(-1)^{k+1}}{\theta^{2k+1}}}{1 + \frac{K_2}{K_1}\frac{(-1)^k}{\theta^{2k}}}.$$

Since $\theta > 1$, as $k \rightarrow \infty$ the contribution from $K_1\theta^k$ grows monotonically while $(-1)^k K_2/\theta^k$ alternates and dies to zero. This accounts for the asymptotic behaviour of $C_k = p_k/q_k$ as a single geometric series. The approximate value of q_k for large k is $K_1\theta^k$. C_k can be expanded as a series by the binomial theorem, with first few terms

$$\left(\theta + \frac{K_2}{K_1}\frac{(-1)^{k+1}}{\theta^{2k+1}}\right)\left(1 - \frac{K_2}{K_1}\frac{(-1)^k}{\theta^{2k}}\right). \quad (10.6)$$

Clearly the limiting value of C_k is indeed θ . The other terms are the error ϵ_k (with sign):

$$\epsilon_k \approx -\frac{K_2}{K_1}\frac{(\theta^2 + 1)}{\theta}\frac{(-1)^k}{\theta^{2k}} = (-1)^k K_1 K_2 \frac{(\theta^2 + 1)}{\theta} \frac{1}{q_k^2}.$$

The coefficients simplify greatly because $K_2/K_1 = -1/\theta^2$ and $K_1K_2 = -1/\sigma^2$.

$$\epsilon_k \approx -\frac{(\theta^2 + 1)}{\theta}\left(\frac{-1}{\theta^2}\right)^{k+1} \quad (10.7a)$$

$$\epsilon_k \approx (-1)^{k-1}\frac{(\theta^2 + 1)}{\sigma^2\theta}\frac{1}{q_k^2} = (-1)^{k-1}\frac{1}{\sigma}\frac{1}{q_k^2} \quad (10.7b)$$

I consider Eq 10.7 to be an important pair of statements. The first expression a) gives ϵ_k as a geometric series with common ratio $R = -1/\theta^2$. In the previous subsection, Eq 10.2 gives the common ratio in the series for Δ_k to be $R = -q_{k-2}/q_k$. Using $q_k \approx K_1\theta^k$, $R = -1/\theta^2$, which evaluates to Eq 10.3. So whilst in §10.1.1 it was shown that the differences Δ_k converge to 0 as a geometric series, here we have the stronger result that the errors ϵ_k also converge as geometric series, and with the same common ratio.

The expression Eq 10.7.b ties this analysis back to §1.7 and looks forwards to §11 on the $1/q_k^2$ dependence of errors. Eq 10.7 a), b) together give us the insight that

the $1/q_k^2$ dependence of ϵ_k is intimately related to its
asymptotic convergence as a geometric series.

To tie up §10.1.1 let us return to the geometric series for Δ_k and determine the parameters \tilde{C}_0 and $\tilde{\Delta}_1$. $\tilde{\Delta}_1$ is found from the approximation

$$\tilde{\Delta}_1 \approx \frac{1}{q_0q_1} = \frac{1}{K_1^2\theta} = \frac{\sigma^2}{\theta^3} = \frac{1}{2}(a^2 + 4)[\sigma(a^2 + 1) - a(a^2 + 3)].$$

\tilde{C}_0 is found using $(1 - R)(\theta - \tilde{C}_0) = \tilde{\Delta}_1$:

$$\tilde{C}_0 = \frac{1}{2}[a(a^2 + 5) - \sigma(a^2 + 1)].$$

Some numerical values are given in Table 21. The value of $R = -3 + 2\sqrt{2} = -1.762747$ for $a = 2$ is that found numerically in Table 20. To illustrate how quickly the error estimate from the asymptotic series approaches the exact error, Table 22 gives some numerical values for $a = 2$.

Table 21: Values of constants in the asymptotic geometric series for $\{a : \underline{a}\}$

Parameter	$a = 1$	$a = 2$	$a = 3$
$\theta = \frac{1}{2}(a + \sigma)$	$\frac{1}{2}(1 + \sqrt{5})$	$1 + \sqrt{2}$	$\frac{1}{2}(3 + \sqrt{13})$
$R = -1/\theta^2$	$\frac{1}{2}(-3 + \sqrt{5})$	$-3 + 2\sqrt{2}$	$\frac{1}{2}(-11 + 3\sqrt{13})$
$\tilde{\Delta}_1 = \sigma^2/\theta^3$	$5\sqrt{5} - 10$	$40\sqrt{2} - 56$	$65\sqrt{13} - 234$
$\tilde{C}_0 = \frac{1}{2}[a(a^2 + 5) - (a^2 + 1)\sigma]$	$3 - \sqrt{5}$	$9 - 5\sqrt{2}$	$21 - 5\sqrt{13}$
$\epsilon_k q_k^2 = 1/\sigma$	$1/\sqrt{5}$	$1/\sqrt{8}$	$1/\sqrt{13}$

Table 22: Actual errors in convergents compared with estimates from asymptotic series, for $\{2 : \underline{2}\} = 2.41421356237$

k	q_k	C_k	ϵ_k	$(-1)^{k-1}/(\sigma q_k^2)$
0	1	2	-4.142E-01	-3.536E-01
1	2	2.5	8.579E-02	8.839E-02
2	5	2.40	-1.421E-02	-1.414E-02
3	12	2.4167	2.453E-03	2.455E-03
4	29	2.41379	-4.205E-04	-4.204E-04
5	70	2.414286	7.215E-05	7.215E-05
6	169	2.4142012	-1.238E-05	-1.238E-05
7	408	2.41421569	2.124E-06	2.124E-06
8	985	2.4142131980	-3.644E-07	-3.644E-07
9	2378	2.41421362489	6.252E-08	6.252E-08
10	5741	2.414213551646	-1.073E-08	-1.073E-08

10.2 Convergence of $\{b : \underline{a}, b\}$, $L = 2$

The next most simple case has two recurring digits $\theta = \{b : \underline{a}, b\} > 0$. From §3.2.2 θ satisfies ¹⁰

$$a\theta^2 - ab\theta - b = 0 \quad \text{with solution} \quad \theta = \frac{1}{2a}[ab + \sqrt{(ab)^2 + 4ab}].$$

As with $\{a : \underline{a}\}$, the asymptotic geometric series can be revealed by elementary means. Observe that

$$\text{for } k \text{ even} \quad p_k = \frac{b}{a} q_{k+1} \quad \text{and} \quad q_k = p_{k-1}, \quad (10.8a)$$

$$\text{for } k \text{ odd} \quad p_k = q_{k+1} \quad \text{and} \quad q_k = \frac{a}{b} p_{k-1}. \quad (10.8b)$$

The equivalent of Eq 10.2 is

$$\text{For } k \text{ even} \quad -\frac{q_{k-2}}{q_k} \rightarrow R' = \frac{-q_{k-2}}{b q_{k-1} + q_{k-2}} = \frac{-1}{b \left(\frac{q_{k-1}}{q_{k-2}}\right) + 1} = \frac{-1}{b \left(\frac{a}{b}\right) \theta + 1}$$

$$\text{For } k \text{ odd} \quad -\frac{q_{k-2}}{q_k} \rightarrow R' = \frac{-q_{k-2}}{a q_{k-1} + q_{k-2}} = \frac{-1}{a \left(\frac{q_{k-1}}{q_{k-2}}\right) + 1} = \frac{-1}{a\theta + 1}$$

¹⁰I have swapped a and b from the notation in §3.2.2.

These limiting ratios R' of consecutive differences are identical; it does not matter whether k is odd or even. Moreover, it has the same form as Eq 10.2,

$$R' = \frac{-1}{a\theta + 1} = -\frac{1}{2}[ab + 2 - \sqrt{(ab)^2 + 4ab}]. \quad (10.9)$$

Note again a point mentioned in §3.2.2. Because R' is symmetrical in a and b , the rates of convergence are the same for the two fractions $\theta_A = \{b : \underline{a}, b\}$ and $\theta_B = \{a : \underline{b}, a\}$ even though their limits are different. In fact θ_A and θ_B are related through

$$\theta_A = \frac{1}{\theta_B - a}, \quad \theta_B = \frac{1}{\theta_A - b}$$

from which $a\theta_A = b\theta_B$.

Following the previous section, we now consider $\{b : \underline{a}, b\}$ using linear difference equations, to find exact expressions for p_k and q_k . The operator and auxiliary equations now must be constructed from a linear combination of three successive recursion relations as follows:

$$\begin{aligned} q_{k+2} - aq_{k+1} - q_k &= 0 \\ q_{k+1} - bq_k - q_{k-1} &= 0 \\ q_k - aq_{k-1} - q_{k-2} &= 0 \end{aligned}$$

Multiply row 2 by a and row 3 by -1 and add:

$$q_{k+2} - (ab + 2)q_k + q_{k-2} = 0.$$

Let \mathcal{D}_2 be the operator which transforms q_k to q_{k+2} . The operator and auxiliary equations are therefore

$$(\mathcal{D}_2^2 - (ab + 2)\mathcal{D}_2 + 1)q_k = 0, \quad \lambda^2 - (ab + 2)\lambda + 1 = 0,$$

linking every *second* q . Unlike $\{a : \underline{a}\}$, here λ and θ obey different equations. The roots are

$$\lambda_{\frac{1}{2}} = \frac{1}{2}[ab + 2 \pm \sqrt{4ab + a^2b^2}] \quad (10.10)$$

and again there is symmetry in a and b . Clearly $\lambda_1\lambda_2 = 1$. R' in Eq 10.9 equals $-\lambda_2$. The operator equations give $\mathcal{D}_2q_k = q_{k+2} = \lambda_{\frac{1}{2}}q_k$ and express geometric series between every second denominator. There are therefore two geometric series, one for the even q_k , the other for the odd ones. However, the analysis leading to Eq 10.9 suggests only one geometric series. This paradox needs to be resolved.

To express the general solution introduce S and T by

$$T = ab + 4 \quad \text{and} \quad S = \sqrt{a^2b^2 + 4ab} = \sqrt{abT}.$$

$$\text{Then} \quad \lambda_{\frac{1}{2}} = \frac{1}{2}[T - 2 \pm S].$$

The initial conditions for q_k are $q_0 = 1$, $q_1 = a$, $q_2 = ab + 1$, $q_3 = a^2b + 2a$. These give the even denominators as

$$\begin{aligned} q_{2m} &= H_1\lambda_1^m + H_2\lambda_2^m \\ \text{where} \quad H_{\frac{1}{2}} &= \frac{1}{2T}[\pm S + T], \end{aligned}$$

and the odd denominators as

$$q_{2m+1} = K_1 \lambda_1^m + K_2 \lambda_2^m$$

where
$$K_2 = \frac{1}{2bT} [\pm(ab+2)S + abT].$$

The initial conditions for p_k are $p_0 = b$, $p_1 = ab + 1$, $p_2 = ab^2 + 2b$, $p_3 = a^2b^2 + 3ab + 1$. These give the even numerators as

$$p_{2m} = L_1 \lambda_1^m + L_2 \lambda_2^m$$

where
$$L_2 = \frac{1}{2aT} [\pm(ab+2)S + abT].$$

and the odd numerators as

$$p_{2m+1} = M_1 \lambda_1^m + M_2 \lambda_2^m$$

where
$$M_2 = \frac{1}{2T} [\pm(ab+3)S + (ab+1)T].$$

Notice that $bK_1 = aL_1$. Notice also that the variables and coefficients with subscript 1, with the + signs, are larger in magnitude than those with subscript 2. Using the binomial expansion as for Eq 10.6 one can show that

$$\frac{\tilde{\Delta}_{k-1}}{\tilde{\Delta}_k} = -\frac{q_{k-1}}{q_k} \rightarrow R' = \frac{-1}{\lambda_1} = -\lambda_2, .$$

$$q_{2m} \rightarrow H_1 \lambda_1^m, \quad q_{2m+1} \rightarrow K_1 \lambda_1^m.$$

$$p_{2m} \rightarrow L_1 \lambda_1^m, \quad p_{2m+1} \rightarrow M_1 \lambda_1^m.$$

Other relations are

$$\theta = \frac{L_1}{H_1} = \frac{M_1}{K_1}, \quad \frac{-b}{a\theta} = \frac{L_2}{H_2} = \frac{M_2}{K_2}, \quad \lambda_1 = a\theta + 1, \quad \lambda_2 = \frac{1}{(a\theta + 1)}$$

The exact expression for convergent C_k depends on whether $k = 2m$ or $k = 2m + 1$:

$$C_{2m} = \frac{L_1 \lambda_1^m + L_2 \lambda_2^m}{H_1 \lambda_1^m + H_2 \lambda_2^m} \approx \left[\theta + \frac{L_2}{H_1} \lambda_2^{2m} \right] \left[1 - \frac{H_2}{H_1} \lambda_2^{2m} \right],$$

$$C_{2m+1} = \frac{M_1 \lambda_1^m + M_2 \lambda_2^m}{K_1 \lambda_1^m + K_2 \lambda_2^m} \approx \left[\theta + \frac{M_2}{K_1} \lambda_2^{2m} \right] \left[1 - \frac{K_2}{K_1} \lambda_2^{2m} \right],$$

This furnishes estimates for the errors ϵ_{2m} , ϵ_{2m+1} :

$$\epsilon_{2m} \approx [L_2 - H_2\theta] \frac{1}{H_1 \lambda_1^{2m}} < 0,$$

$$\epsilon_{2m+1} \approx [M_2 - K_2\theta] \frac{1}{K_1 \lambda_1^{2m}} > 0,$$

where I have again used $\lambda_2 = 1/\lambda_1$. Immediately we recognise $H_1^2 \lambda_1^{2m} = q_{2m}^2$ and $K_1^2 \lambda_1^{2m} = q_{2m+1}^2$, so the ubiquitous $1/q_k^2$ dependence of error is found again. The coefficients simplify, giving the signed errors as

$$\epsilon_{2m} \approx -\frac{S}{aT} \frac{1}{q_{2m}^2} = -\sqrt{\frac{b}{a}} \frac{1}{\sqrt{ab+4}} \frac{1}{q_{2m}^2}, \quad (10.11a)$$

$$\epsilon_{2m+1} \approx \frac{S}{bT} \frac{1}{q_{2m+1}^2} = \sqrt{\frac{a}{b}} \frac{1}{\sqrt{ab+4}} \frac{1}{q_{2m+1}^2}. \quad (10.11b)$$

Both reduce to Eq 10.8 when $b = a$. These formulae have different coefficients, weighted by the ratio b/a or a/b respectively. However

$$\frac{\epsilon_{2m+1}}{\epsilon_{2m}} = \left(\frac{M2 - K_2\theta}{L_2 - H_2\theta} \right) \frac{H_1}{K_1} = R' = -\lambda_2.$$

This allows us to write a single formula for the signed error in terms of λ_2 (though not $1/q_k$):

$$\epsilon_k \approx (-1)^k \frac{S}{a} \lambda_2^{k+1} = \frac{\sqrt{a^2b^2 + 4ab}}{a} (-\lambda_2)^{k+1} \quad (10.12)$$

for all values of k , with λ_2 given by Eq 10.10.

This means that, although there are two interleaving asymptotic geometric series, not only do they share the same common ratio, but they have no offset in first terms, so they coincide. It also suggests that λ_2 is a more profound parameter than the denominators q_k , even though our analysis so far, and much of the literature, focuses on comparison of error with powers of $1/q_k$.

One last observation before leaving the $L = 2$ case: observe that Eqs 10.11 a, b can be rearranged to read

$$|a \epsilon_{2m} q_{2m}^2| = |b \epsilon_{2m+1} q_{2m+1}^2| = \frac{S}{T} = \sqrt{\frac{ab}{ab+4}} \quad (10.12).$$

Since $\{b : \underline{a}, b\}$ has b as the even partial quotients and a as the odd ones, these two expressions have for all k the form that $a_{k+1} \epsilon_k q_k^2$ is constant. This seems to hold at least approximately for more general continued fractions; some numerical evidence for this is presented in §11.4

10.3 Fractions with recurring sequence length $L \geq 3$

So far in §10 we have seen that continued fractions of the two forms $\{a : \underline{a}\}$ and $\{b : \underline{a}, b\}$ have differences errors ϵ_k which tend to zero asymptotically as single geometric progressions, and behave such that $a_{k+1} \epsilon_k q_k^2$ is constant. When there are three or more recurring partial quotients, the behaviour changes and the recurring sequence, length L , imposes a periodic variation with period L . The limiting pattern in the convergents of θ as $k \rightarrow \infty$ is of L interwoven geometric series, all with the same common ratio but with L different starting values.

To illustrate typical behaviour, take the $L = 4$ case of $\theta = \{0 : \underline{1}, \underline{2}, \underline{1}, \underline{6}\} = \sqrt{14} - 3$. Table 23 below corresponds to Table 21 for $\{2 : \underline{2}\}$. Note that the differences in the last column are periodic, repeating every fourth convergent. The horizontal lines divide the various periods. Figure 8 is a graph of the differences. Close examination will convince the reader that, superimposed on the straight log-linear trend line, is a periodic variation with period 4. Figure 9 plots the second differences to bring out this cyclic pattern.

Accordingly we calculate the gradient of the log-linear plot through every *fourth* value of k . From the graph we can expect these all to be the same and about equal to -1.7012 , as read from the fitted line in Figure 8. Now

$$\frac{\Delta_k}{\Delta_{k-4}} = \frac{q_{k-4} q_{k-5}}{q_k q_{k-1}}.$$

From the high convergents C_k of Table 24 the value of this is $-6 \cdot 80017$ for any value of k , so the gradient of the graph is $-6 \cdot 80017/4 = -1 \cdot 70004$. $\exp(-1 \cdot 70004) = 0 \cdot 1827$. Moreover, the intercept on the fitted trend line in Figure 8 gives the constant $e^{0.53} \approx 1 \cdot 7$. Therefore *on average* the differences Δ_k tend to zero as a quasi-geometric series $1 \cdot 7 \times 0 \cdot 1827^k$, though with a ripple of period 4. Shortly I will derive the exact value of this.

Table 23: Logarithm of differences between successive convergents of $\{0 : 1, 2, 1, 6\}$

k	C_k	$-\ln \Delta_k $	$\frac{\ln \Delta_k - \ln \Delta_{k-1} }{= \ln \Delta_k / \Delta_{k-1} }$
0	1/1		
1	2/3	-1.09861229	
2	3/4	-2.48490665	-1.38629436
3	20/27	-4.68213123	-2.19722458
4	23/31	-6.72982407	-2.04769284
5	66/89	-7.92262357	-1.19279950
6	89/120	-9.27612811	-1.35350454
7	600/809	-11.48329066	-2.20716255
8	689/929	-13.52990766	-2.04661700
9	1978/2667	-14.72281826	-1.19291061
10	2667/3596	-16.07628692	-1.35346866
11	17980/24243	-18.28346059	-2.20717367
12	20647/27839	-20.33007639	-2.04661580
13	59274/79921	-21.52298712	-1.19291073
14	79921/107760	-22.87645574	-1.35346862
15	538800/726481	-25.08362942	-2.20717368
16	618721/834241	-27.13024522	-2.04661580
17	1776242/2394963	-28.32319169	-1.19294648
18	2394963/3229204	-29.67662457	-1.35343287
19	16146020/21770187	-31.88379825	-2.20717368
20	18540983/24999391	-33.93041405	-2.04661580
21	53227986/71768969	-35.12332478	-1.19291073
22	71768969/96768360	-36.47679339	-1.35346862
23	483841800/652379129	-38.68396708	-2.20717368

As with the earlier cases in §10.1, 10.2, it is possible to obtain an auxiliary equation and thence explicit, though cumbersome, expressions for the various p_k, q_k . One finds that every L^{th} numerator or denominator is linked by the auxiliary equation. The auxiliary equation is formed from a linear combination of $2L - 1$ successive recursion relations. I will illustrate the method for $L = 4$ and the continued fraction $\{d : a, b, c, d\}$. The form of a recursion relation is $p_{k+1} - ap_k - p_{k-1} = 0$, $q_{k+1} - aq_k - q_{k-1} = 0$. Draw up an array of coefficients of 7 such successive equations, and multiply the rows by suitable constants so that the sums of the second, third, fifth and sixth columns are each zero. The header denotes the index of p or q with $+2$ meaning p_{k+2} or q_{k+2}

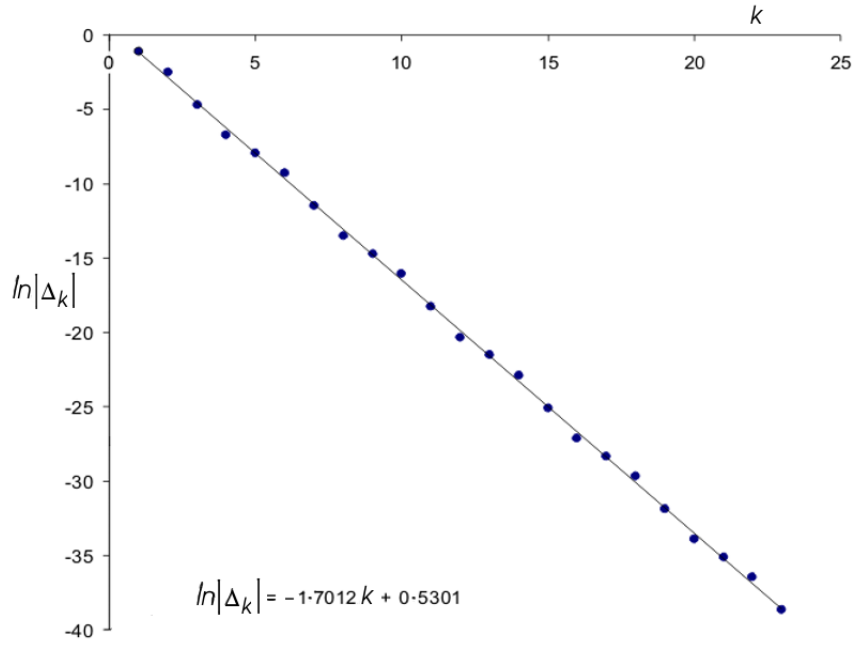


Figure 8: Logarithm of differences $|\Delta_k|$ between successive convergents of $\{0 : \underline{1, 2, 1, 6}\}$

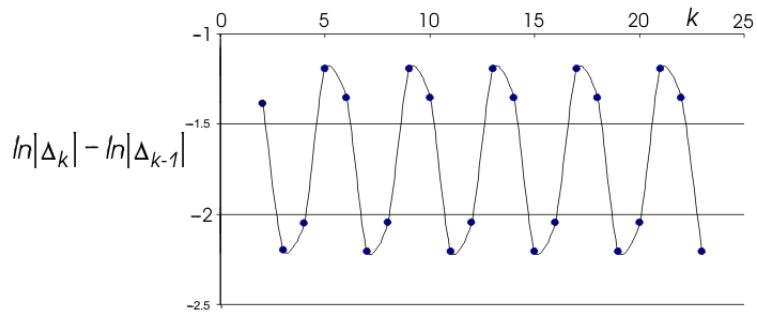


Figure 9: Second differences $\ln|\Delta_k| - \ln|\Delta_{k-1}|$ of $\{0 : \underline{1, 2, 1, 6}\}$

4	3	2	1	0	-1	-2	-3	-4
1	-d	-1						
	1	-c	-1					
		1	-b	-1				
			1	-a	-1			
				1	-d	-1		
					1	-c	-1	
						1	-b	-1

The modified array is

$$\begin{array}{cccccccccccc}
4 & 3 & 2 & & 1 & & 0 & & -1 & & -2 & -3 & -4 \\
\hline
1 & -d & -1 & & & & & & & & & & & \\
& d & -cd & & -d & & & & & & & & & \\
& & cd+1 & -(cd+1)b & -(cd+1) & & & & & & & & & \\
& & & bcd+d+b & -a(bcd+d+b) & -(bcd+d+b) & & & & & & & & \\
& & & & -bc-1 & d(bc+1) & bc+1 & & & & & & & \\
& & & & & b & -bc & -b & & & & & & \\
& & & & & & -1 & b & 1 & & & & &
\end{array}$$

Adding the rows now gives the auxiliary equation linking every fourth numerator

$$\lambda^2 - (abcd + ab + bc + cd + da + 2)\lambda + 1 = 0.$$

This has cyclic symmetry over a, b, c, d , confirming that all of the four interwoven limiting geometric series have the same common ratio.

Using a symbolic matrix algebra program I have evaluated the auxiliary equation for recursion lengths up to $L = 8$. For L odd, the form is $\lambda^{2L} - \rho_L \lambda^L - 1$, and for L even it is $\lambda^{2L} - \rho_L \lambda^L + 1$, where ρ_L is the coefficient of λ . Thus

$$L = 2: \quad \lambda^4 - (ab + 2)\lambda^2 + 1,$$

$$L = 3: \quad \lambda^6 - (abc + a + b + c)\lambda^3 - 1,$$

$$L = 4: \quad \lambda^8 - (abcd + ab + bc + cd + da + 2)\lambda^4 + 1.$$

The solutions for λ^L are clearly

$$L \text{ even:} \quad \lambda_{\frac{1}{2}}^L = \frac{1}{2}(\rho_L \pm \sqrt{\rho^2 - 4}),$$

$$L \text{ odd:} \quad \lambda_{\frac{1}{2}}^L = \frac{1}{2}(\rho_L \pm \sqrt{\rho^2 + 4}),$$

Similar to the analysis of §10.2, the common ratio \mathcal{R}_L of the series for every L^{th} error is the remarkably simple value

$$\mathcal{R}_L = \frac{\epsilon_{k+L}}{\epsilon_k} = \begin{cases} +\lambda_2^2 & \text{if } L \text{ is even} \\ -\lambda_2^2 & \text{if } L \text{ is odd} \end{cases} \quad (10.14)$$

(Beware: \mathcal{R}_L relates every L^{th} convergent, not to be confused with R from §10.1 or R' from §10.2 which relate *adjacent* convergents.)

The expressions for ρ_L are listed below, each arranged to emphasise the various constituent sets of letter permutations:

$$\begin{aligned}
\rho_2 &= ab + 2 \\
\rho_3 &= abc + a + b + c \\
\rho_4 &= abcd + ab + bc + cd + da + 2 \\
\rho_5 &= abcde + abc + bcd + cde + dea + eab + a + b + c + d + e
\end{aligned}$$

$$\begin{aligned} \rho_6 = & abcdef + abcd + bcde + cdef + defa + efab + fabc \\ & + ab + bc + cd + de + ef + fa + ad + be + cf + 2 \end{aligned}$$

$$\begin{aligned} \rho_7 = & abcdefg \\ & + abcde + bcdef + cdefg + defga + efgab + fgabc + gabcd \\ & + abc + bcd + cde + def + efg + fga + gab \\ & + abe + bcf + cdg + dea + efb + fgc + gad \\ & + a + b + c + d + e + f + g \end{aligned}$$

$$\begin{aligned} \rho_8 = & abcdefgh \\ & + abcdef + bcdefg + cdefgh + defgha + efghab + fghabc + ghabcd + habcde \\ & + abcd + bcde + cdef + defg + efg + fgha + gh + habc \\ & + abcf + bcdg + cdeh + defa + efgb + fg + ghc + ghad + habe \\ & + abef + bcfg + cdgh + deha \\ & + ab + bc + cd + de + ef + fg + gh + ha \\ & + ad + be + cf + dg + eh + fa + gb + hc \\ & + 2. \end{aligned}$$

All ρ_L are invariant under cyclic permutation of the letters. Because of this, for any L all L interwoven geometric series within one continued fraction have the same common ratio \mathcal{R}_L . Each ρ_L is made of sets of monomials with $L, L-2, L-4, \dots$ letters. All the cyclic permutations of $L-2, L-4$, etc. letter occur, plus, for $L \geq 6$ additional non-cyclic permutations. To see what is happening here, look at ρ_7 and picture the letters a to g written in a clockwise circle (Figure 10). The letters lie at the vertices of a 7-gon. The monomials which make up ρ_7 are those which, as a set, are invariable under the symmetry actions of a regular 7-gon, which form the dihedral group D_7 ¹¹. For example, for ρ_7 Figure 10 shows one of the lines of mirror symmetry, and ρ_7 is invariant under reflection about this line, which interchanges the pairs $(b, g), (c, f), (d, e)$, leaving a fixed. In contrast, exchanging only, say, (b, g) would create monomials abf and $abdef$ which are not present in ρ_7 .

Higher values of L produce richer symmetries and hence more terms in ρ_L . Figure 11 illustrates some monomials in ρ_8 . Rows 1, 2, 3 and 6 of ρ_8 are cyclic permutations with 8, 6, 4 and 2 adjacent letters respectively in each. Row 4 involves 3 adjacent letters plus the letter opposite the central one: *e.g.* f is opposite to b in Figure 10a. Row 4 combines two opposite pairs such as ab and ef . The eight pairs in the penultimate row, when connected pairwise, form the 8-pointed star for Figure 10b.

However, not all invariant combinations of letters are included in the make-up of ρ_L . For instance, with $L = 5$ the combination set $\{acd, bde, cea, dab, ebc\}$ is not present even though it is invariant under D_5 . Similarly, from Figure 10 we might expect both bhd and bge to be in ρ_7 , but they are not. I doubt, therefore, whether it would be possible to determine ρ_L combinatorially, by selecting from the combinations of $L, L-2$, etc. letters those which do not vary under action of D_7 .

¹¹There is an alternative notation for this group as D_{14} , where the 14 counts the number of symmetry operations, not the number of vertices.

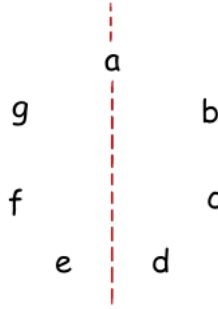


Figure 10: Invariance of ρ_7 under dihedral D_7 symmetry.

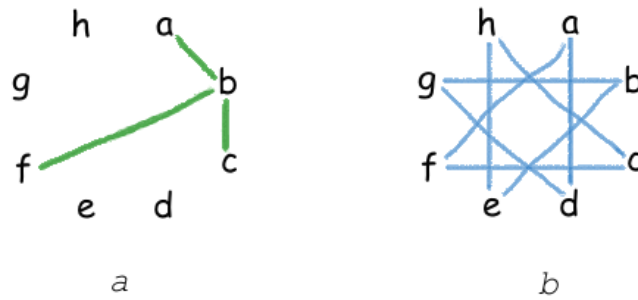


Figure 11: Illustrating symmetries in rows 4 and 7 of the formula for coefficient ρ_8 .

For any fixed values of the L letters, there will be $L!$ permutations of the recursion sequence of partial quotients. These will fall into sets which are equivalent in the sense of having the same rate of convergence, i.e. the same \mathcal{R}_L . Each such set will contain those $2L$ permutations which are unchanged under action of D_L . Hence we can expect $L!/(2L) = (L-1)!/2$ different rates of convergence within the $L!$ permutations. For $L = 2$ this evaluates to 1, consistent with the single geometric series in §10.2 and its invariance to exchange of a and b . For $L = 3$ this also predicts only one common ratio, and some numerical investigation bears this out. For $L = 4$ three distinct values of \mathcal{R}_3 are predicted and this too is supported by my numerical evidence from trial cases using a spreadsheet.

There is another intriguing feature of the ρ_L . The number of terms in ρ_L increases as 2, 4, 6, 11, 17, 29, 46 for $L = 2$ to 8 respectively. The comprehensive ‘On-Line Encyclopedia of Integer Sequences’ suggests that these may be related either to powers of the Golden Ratio, or to products of two Fibonacci numbers. The next number, for ρ_9 , would be 76 or 77. Scope for further investigation!

10.3.1 Specific case of $L = 3$

Explicit formulae for errors in the convergents, $\epsilon_k = |C_k - \theta|$, rapidly become complicated for all $L > 2$. I have, however, evaluated the case $L = 3$, mainly to see if the relationship $a_{k+1} \epsilon_k q_k^2 \approx \text{constant}$ appears still to hold. $\mathcal{S} = \sqrt{\rho_3^2 + 4}$ takes the place of σ from §10.1 or S from §10.2. For $\{c : \underline{a, b, c}\}$

$$\mathcal{S} = \sqrt{(abc + a + b + c)^2 + 4}.$$

The continued fraction has value

$$\theta = \frac{\rho - 2a + \mathcal{S}}{2(ab + 1)}$$

where ρ means ρ_3 . The parameters λ_1, λ_2 are

$$\lambda_{1,2} = \frac{1}{2}(\rho \pm \mathcal{S}),$$

from which the common ratio of the asymptotic geometric series, spanning every third convergent, is $R = \lambda_2^2$. Accordingly the average rate of convergence, in the sense of the geometric mean of $\epsilon_k/\epsilon_{k-1}$, must be $\lambda_2^{2/3}$. Approximate values for the denominators $q_{3m}, q_{3m+1}, q_{3m+2}$ are

$$q_{3m} = H_1 \lambda_1^m, \quad q_{3m+1} = J_1 \lambda_1^m, \quad q_{3m+2} = K_1 \lambda_1^m \quad (10.15)$$

where the coefficients are

$$H_1 = \frac{1}{2\mathcal{S}^2}[\mathcal{S}^2 + \rho\mathcal{S} - 2b\mathcal{S}], \quad J_1 = \frac{1}{2\mathcal{S}^2}[a(\mathcal{S}^2 + \rho\mathcal{S}) + 2\mathcal{S}], \quad K_1 = \frac{1}{2\mathcal{S}^2}(ab + 1)(\mathcal{S}^2 + \rho\mathcal{S}).$$

(Note that $\mathcal{S}^2 = \rho^2 + 4$.) Here again is the $1/q_k^2$ dependence of error. The equivalents of Eqs 9.12 a, b have the cyclic symmetric forms

$$\epsilon_{3m} = -\frac{bc + 1}{\mathcal{S}} \frac{1}{q_{3m}^2}, \quad \epsilon_{3m+1} = \frac{ca + 1}{\mathcal{S}} \frac{1}{q_{3m+1}^2}, \quad \epsilon_{3m+2} = -\frac{ab + 1}{\mathcal{S}} \frac{1}{q_{3m+2}^2}. \quad (10.16)$$

Relations of exactly the same form to Eq 10.16 and 10.17 will occur for $L > 3$, meaning that $\epsilon_k \propto 1/q_k^2$ is an essential characteristic of quadratic irrationals. Finally, noting that the next partial quotients a_{k+1} for $k \equiv 0 \pmod{3}, k \equiv 1, k \equiv 2$ are respectively a, b, c , the ‘ $a_{k+1} \epsilon_k q_k^2 \approx \text{constant}$ ’ conjecture reads as

$$a \epsilon_{3m} q_{3m}^2 = \frac{\rho - b - c}{\mathcal{S}}, \quad b \epsilon_{3m+1} q_{3m+1}^2 = \frac{\rho - c - a}{\mathcal{S}}, \quad c \epsilon_{3m+2} q_{3m+2}^2 = \frac{\rho - a - b}{\mathcal{S}}.$$

Clearly these are not equal constants, so illustrating a limitation to the conjecture. However, if a, b, c are not too small and not greatly different from each other, the relation will be approximately correct.

10.4 The transition from $\{1 : \underline{1}\}$ to $\{2 : \underline{2}\}$

The continued fraction $\{1 : \underline{1}\}$ is a limiting case, being the slowest of all to converge. This section presents some numerical results, mainly graphically, of the patterns in the convergents as partial quotients ‘2’ are gradually introduced into the recurring continued fraction $\{1 : \underline{1}\}$ until all ‘1’ have been replaced by ‘2’. This numerical study gives us some appreciation of what happens when the 1s are diluted. These results are make an interesting pattern in their own right, and are used in §11.

Using an arbitrary precision software package I have calculated the first 41 convergents C_k and errors ϵ_k for various recurring sequence made of digits 1 and 2. C_{41} has been taken as a sufficient approximation to the limiting real had the recurring sequence continued to infinity. Various statistics could be calculated, but I have concentrated on the product $\epsilon_k q_k^2$ because, as proved in §10.1, 10.2, it is a characteristic constant when $L = 1, 2$ and may be approximately constant in other cases.

Figure 12 presents $\epsilon_k q_k^2$ for $a_{16}, a_{33} = 2$, all other $a_k = 1$. You may see an analogy with wave motion in which a single impulse of ‘2’ is struck against a steady background of 1s. The fraction $\{1 : \underline{1}\}$ has the constant value of $\epsilon_k q_k^2 = 1/\sqrt{5} = 0.4472$. At the other limit of $\{2 : \underline{2}\}$ the value would

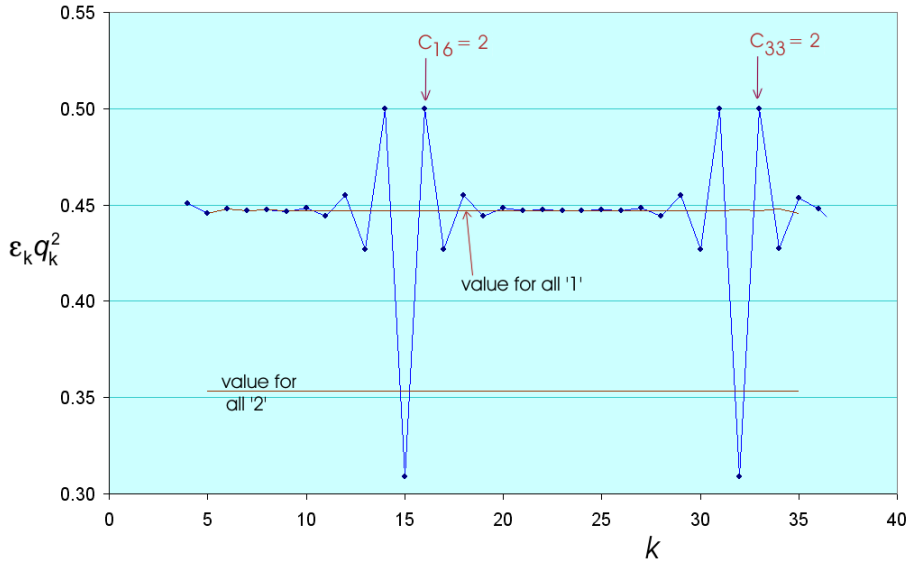


Figure 12: Effect on $\epsilon_k q_k^2$ of introducing a single 2 into $\{1 : \underline{1}\}$ at positions 16 and 33

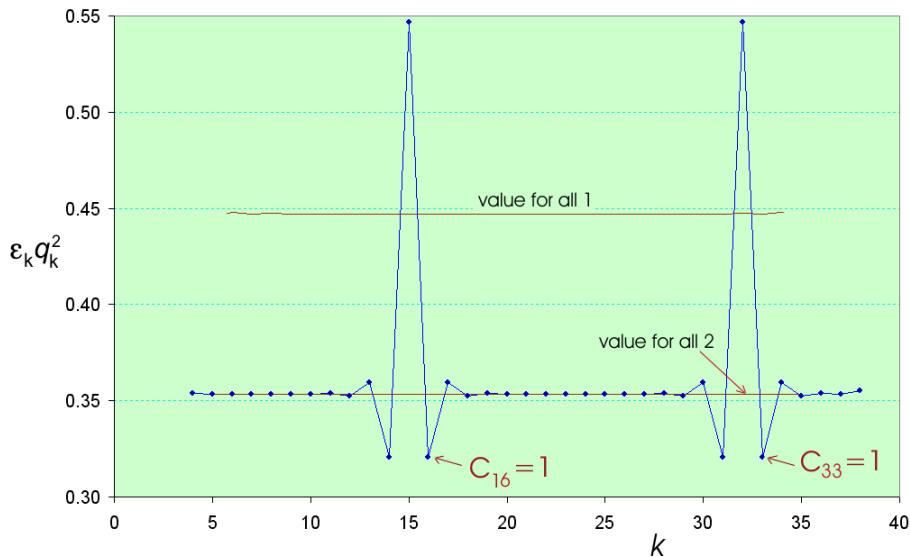


Figure 13: Complementary case to Figure 12, with single 1 introduced into $\{2 : \underline{2}\}$ at positions 16 and 33

be $1/\sqrt{8} = 0.3536$, also shown. The effect of adding the single 2 is to cause a ‘splash’ – a trough at $k = 15$ (and at 32), with ‘ripples’ attenuating almost symmetrically to either side.

It is remarkable that the trough occurs at k one less than the position of the 2. It is as if the convergents can anticipate the occurrence of the 2. We saw a similar phenomenon in §6 on Pell’s equation, where the lowest error κ was at positions $k = nL - 1$, and other values were symmetrically placed either side of this. The depth of the trough caused by an isolated 2 can be determined by

applying the method of Eq 6.3. Let $\theta = \{0 : \underline{1}, 1, 1, 1, \dots, 1, 1, 2\}$ and consider the error in C_{L-1} .

$$\theta = \frac{(2 + \theta)p_{L-1} + p_{L-2}}{(2 + \theta)q_{L-1} + q_{L-2}},$$

$$\epsilon_{L-1} = \frac{1}{q_{L-1}[(\theta + 2)q_{L-1} + q_{L-2}]},$$

$$\epsilon_{L-1} q_{L-1}^2 = \frac{1}{\theta + 2 + \frac{q_{L-2}}{q_{L-1}}}.$$

Now the single 2 is so diluted by the 1s that $\theta \approx 1/G$ and $q_{L-2}/q_{L-1} \approx 1/G$ also, G being the Golden Mean. Then

$$\epsilon_{L-1} q_{L-1}^2 \approx \frac{1}{\frac{2}{G} + 2} = \frac{\sqrt{5} - 1}{4} = 0.30902,$$

consistent with Figure 12.

If you look back to §2.4, Eq 2.3, you will recall that the above equation for error relative to q_k^2 was expressed in terms of two continued fractions, $\theta_{k+1} + \chi_k$, formed by splitting the sequence of partial quotients for θ after a_k . The largest sum $\theta_{k+1} + \chi_k$ which can be formed using a single partial quotient 2 in a uniform sea of 1s is $\{2 : \underline{1}\} + \{0 : \underline{1}\} = 1 + G + 1/G = 1 + \sqrt{5}$, giving relative error 0.30902 as above. Similarly, the smallest sum is $\{1 : \underline{1}\} + \{0 : 2, \underline{1}\} = G + 1/(1 + G) = 2$, giving $\epsilon_k q_k^2 = 0.5$ as in Figure 12.

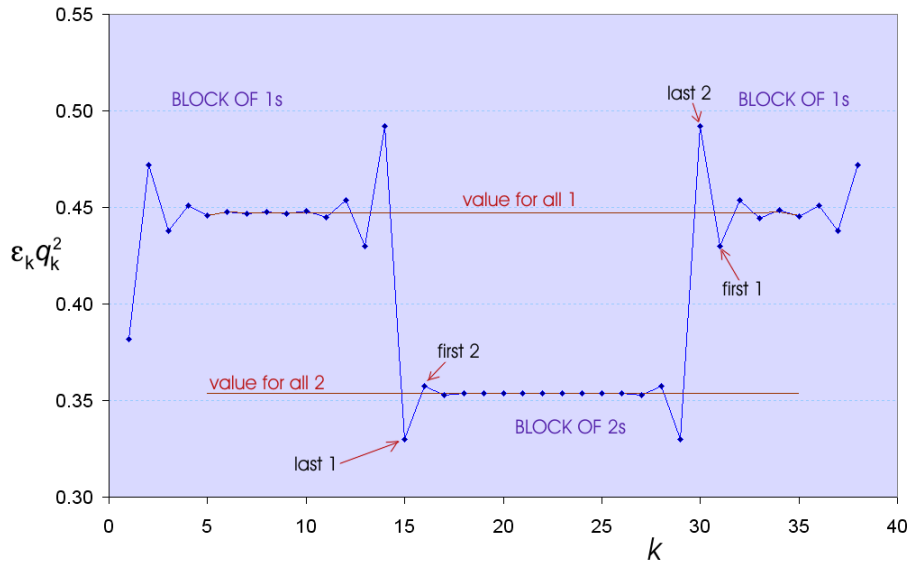


Figure 14: Values of $\epsilon_k q_k^2$ against k , where a_k sequence has blocks of 1 then blocks of 2, each 15 partial quotients long

Figure 13 is the complement of Figure 12, in which a single 1 has been introduced into $\{2 : \underline{2}\} = 1 + \sqrt{2}$. The behaviour is similar though complementary, with a peak now at position $k = nL - 1$. Similar analysis to the above gives the height of the peak as $\frac{1}{7}(1 + 2\sqrt{2}) = 0.54692$.

Another type of impulse is the step function. Figure 14 shows the effect of switching from a sequence of 1s to 2s, and then back. Note the positions in k at which the partial quotients switch from

1 to 2, then back to 1. The overshooting at the ends of the blocks of 1 or 2 is reminiscent of Gibb's phenomenon when finite Fourier series approximate a step discontinuity in a periodic function. As in Figures 11 and 12, away from the discontinuity, the values of $\epsilon_k q_k^2$ remain close to the limiting values of $\{1 : \underline{1}\}$ and $\{2 : \underline{2}\}$.

Further illustration of the transition from $\{1 : \underline{1}\}$ to $\{1 : \underline{2}, 1\}$ is shown in the composite panels of Figures 14 and 15. The vertical axes are $\epsilon_k q_k^2$ values. The notation $1_8 2$ means that there are eight 1s in the recursion sequence then one 2: that is, 11111112, with $L = 9$. The position of the single 2 is marked by a red dot with the digit 2 by it. Observe how, as $\{1 : \underline{2}, 1\}$ is approached, the maximum and minimum values of $\epsilon_k q_k^2$ depart from $1/\sqrt{5}$ and $1/\sqrt{8}$, moving to $1/\sqrt{3} = 0.57735$ and $1/\sqrt{12} = \frac{1}{2}1/\sqrt{3} = 0.2887$ at $\{1 : \underline{2}, 1\}$, consistent with Eq 10.12. It is interesting to obtain these limits using the two series of partial quotients, $\theta_{k+1} + \chi_k$ from Eq 2.3. The largest such sum made by splitting the alternating pattern ...121212... is $\{2 : \underline{1}, 2\} + \{0 : \underline{1}, 2\} = (1 + \sqrt{3}) + (\sqrt{3} - 1) = 2\sqrt{3}$. The smallest sum is $\{1 : \underline{2}, 1\} + \{0 : \underline{2}, 1\} = \sqrt{3}$.

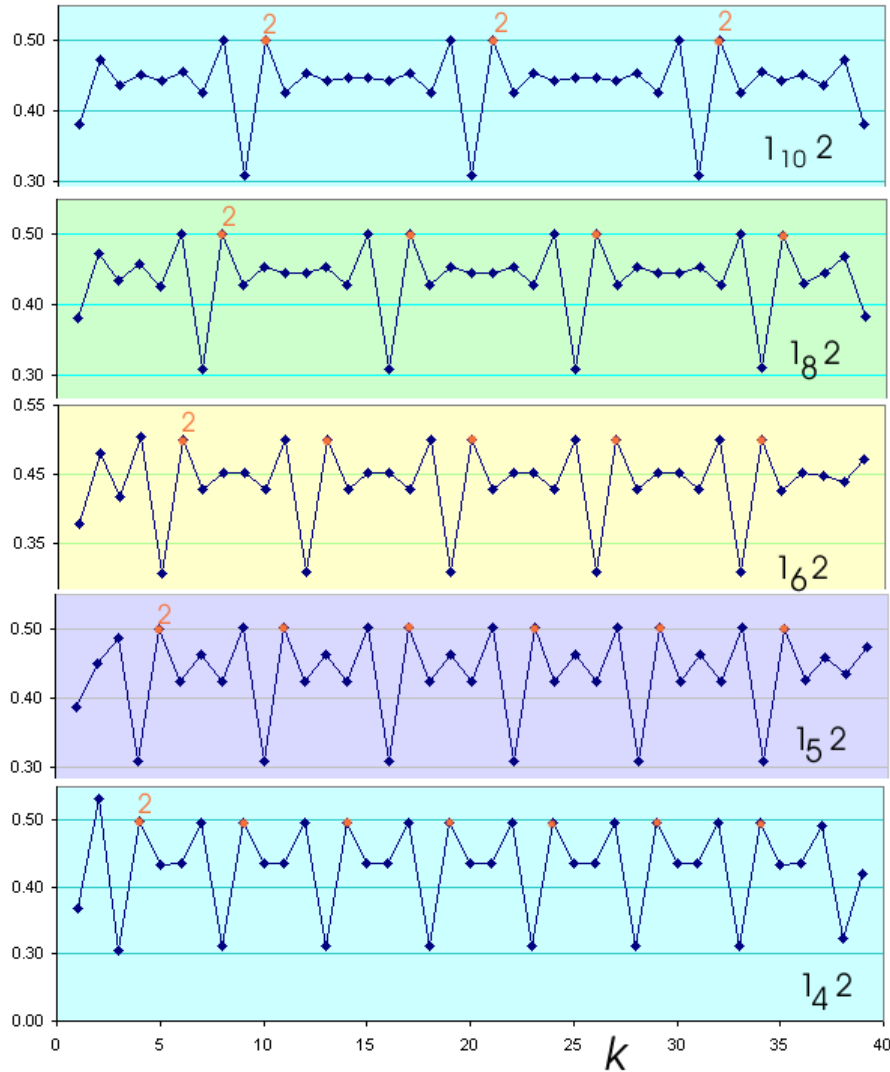


Figure 15: Panels showing $\epsilon_k q_k^2$ versus k in the transition towards $\{1 : \underline{2}, 1\}$.

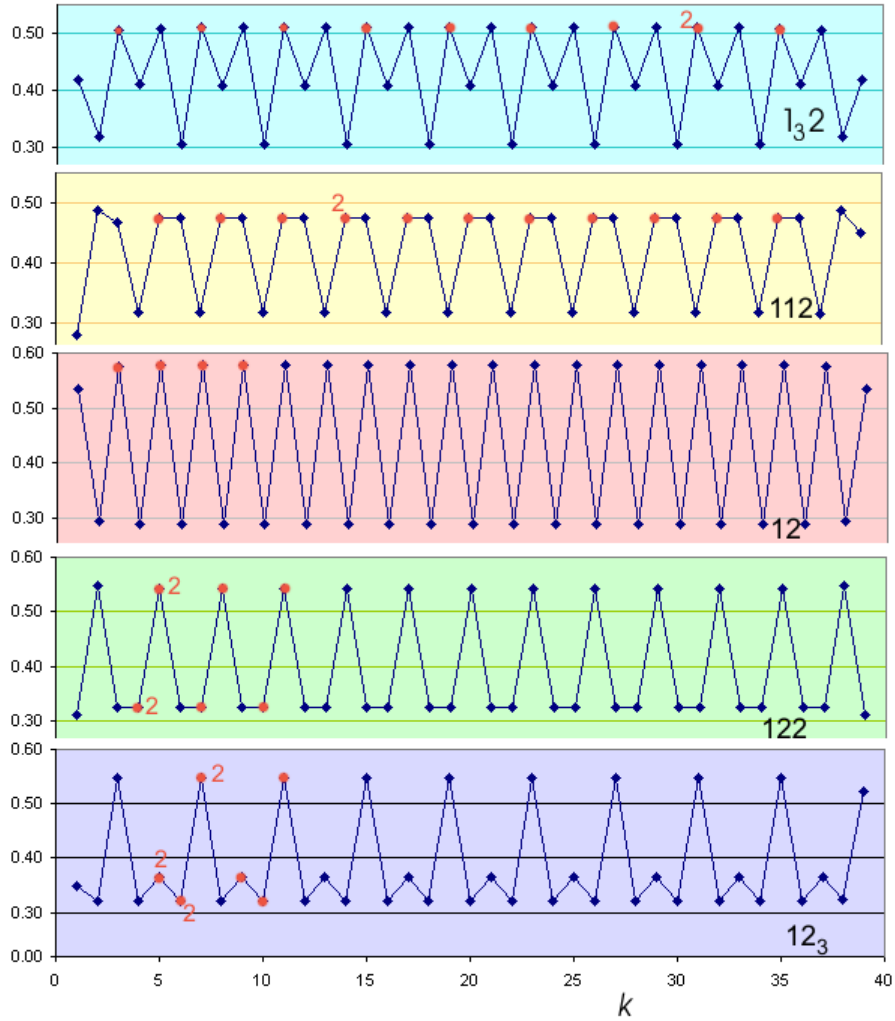


Figure 16: Follow-on panels from Figure 15 showing $\epsilon_k q_k^2$ in the transition through $\{1 : \underline{2}, 1\}$.

Another interesting set of sequences, particularly relevant to a later section, §11.4, is that in which the 1 and 2 partial quotients appear as multiples of the doubles 11 and 22. Figure 17 shows three panels for the convergents of continued fractions with recursion sequences $1_4 2_2$, $1_2 2_2$ and $1_2 2_4$ respectively. On each panel are marked in green the convergents for which the last partial quotient a_k is 1 or 2. The lowest value of $\epsilon_k q_k^2$ occurs at the second '1'. In each panel this minimum value is close to $\frac{1}{3}$.

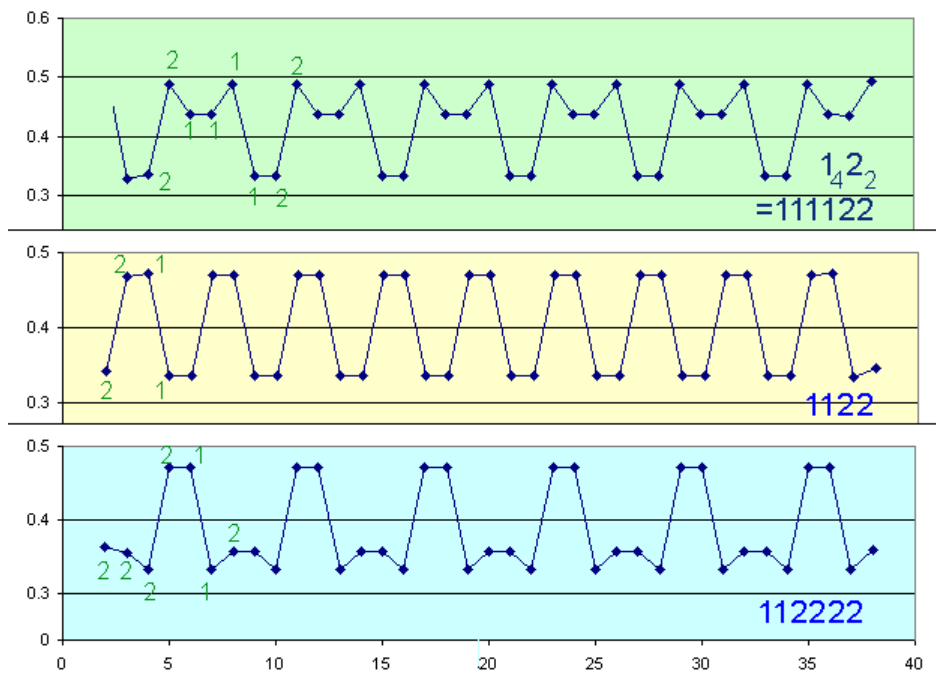


Figure 17: Panels showing $\epsilon_k q_k^2$ versus k for convergents of $\{0 : \underline{1_4 2_2}\}$, $\{0 : \underline{1_2 2_2}\}$ and $\{0 : \underline{1_2 2_4}\}$

11 Further evidence on bounds to errors ϵ_k

11.1 Introductory concepts in Diophantine approximation

Much of this article so far has been looking at the error in approximating a real number θ with any of the convergents C_k of its continued fraction expansion. §1.7, §9.2 and §10 have analysed the difference $\epsilon_k = |C_k - \theta|$, focusing on quantifying this approximation – for instance, knowing that $\pi \approx 22/7$, with error about 0.1%. In this section and the next two (§12, §13) we redirect the focus onto the real number θ itself, asking what can be learned about the structure of \mathbb{R} from approximation by continued fractions. Because the convergents are the most effective of all rational approximations, this section will in effect be discussing in general the approximation of $\theta \in \mathbb{R}$ by $u/v \in \mathbb{Q}$. We will examine the ‘willingness’ of any $\theta \in \mathbb{R}$ to be approximated by rationals; that is, at whether the number of fractions u/v which approximate θ to an accuracy greater than some specified tight limit is finite or infinite, and whether the denominators v must be large integers to achieve this accuracy.

Four main facts have so far been established. For all $\theta \in \mathbb{R}$:

1. $\epsilon_k < \frac{1}{q_k^2}$,
2. More specifically, $\frac{1}{(a_{k+1} + 2)q_k^2} < \epsilon_k < \frac{1}{a_{k+1}q_k^2}$,
3. $\epsilon_k + \epsilon_{k+1} < \frac{1}{2q_k^2} + \frac{1}{2q_{k+1}^2}$, meaning that no two adjacent convergents can both have errors greater than $1/(2q_k^2)$. This was proved in §1.7.
4. For some square roots and other quadratic irrationals, $\epsilon_k q_k^2$ is constant. This was shown in §10.

The relation $\epsilon \propto 1/q^2$ occurs time and time again. We will therefore examine its generalisation

$$\epsilon = \left| \frac{u}{v} - \theta \right| < \text{ or } > \frac{1}{\mathcal{A}v^\alpha},$$

or, in terms of the convergents of θ ,

$$\epsilon_k = |C_k - \theta| < \text{ or } > \frac{1}{\mathcal{A}q_k^\alpha}. \quad (11.1)$$

The analysis will examine three features of these relations:

1. values of the coefficient \mathcal{A} , which will generally depend on θ ,
2. the exponent α , which need not be an integer,
3. the sense of the inequality sign.

Essentially there are two ways of classifying reals, using \mathcal{A} and α respectively. The subject is ‘Diophantine approximation’ and has a literature extending back to the nineteenth century.

Regarding the coefficient \mathcal{A} , it will be obvious that the larger \mathcal{A} becomes, the narrower and tighter becomes the required limit to error. We therefore might expect fewer convergents, and hence fewer rationals of any denominator, to have errors less than this limit. It turns out that, for any given value of $\mathcal{A} > 1$, some reals θ have an infinity of convergents which squeeze under this limit,

whilst other reals have none or only a finite number. In this way \mathcal{A} acts as a sieve on \mathbb{R} , sorting them into categories.

To illustrate the concept of ‘willingness to be approximated’, consider the simplest case of ordinary fractions – the rational numbers \mathbb{Q} . In §1.2 all the convergents of $\theta = \{1 : 2, 3, 4, 5, 6\} = 1393/972$ were listed; there are only five of them, the last being $C_F = \theta$ itself. There are other fractions very close to $1393/972$ – $2787/1945$ is one – but they all have larger denominators, so are considered more complicated, less efficient. Indeed, it is more sensible to say that $1393/972$ is an approximation to $2787/1945$ rather than the other way round. Clearly, with only five close approximations, $\{1: 2, 3, 4, 5, 6\}$ is very resistant to being approximated efficiently by rational numbers.

The essential reason why the rationals are resistant to plentiful approximation by $u/v \in \mathbb{Q}$ is that fractions whose denominators are adjacent integers are quite widely separated on the number line. Think of $1/4$ and $1/5$. The Farey sequence discussion in §8.2 proved that there is no fraction between these with smaller denominator. So if $\theta = p_F/q_F$, all other fractions with denominators close to q_F are a considerable distance away on the number line.

This Section, therefore, introduces some aspects of Diophantine approximation, but there are two large topics which require a Section each to deal with, even at this elementary and introductory level. These are the Lagrange-Markoff Spectrum in §12 which explores \mathcal{A} , and Liouville’s theorem on transcendental numbers in §13 which explores α . Part III closes with §14 regarding Cantor’s proof of transcendental numbers using continued fractions.

11.2 Quadratic forms and Hermite’s $\sqrt{3}$ error limit

This subsection begins our examination of the coefficient \mathcal{A} in Eq 11.1. The analysis will not give a powerfully tighter bound on errors of convergents, but it will develop something of the deep link between continued fractions and the binary quadratic form $ax^2 + 2bxy + cy^2$, which was touched upon at the end of §6 on Pell’s equation.

A ‘binary quadratic form’ is a polynomial in two variables, x and y , in which each term has degree 2. $ax^2 + 2bxy + cy^2$ is abbreviated to $f(x, y) = (a, b, c)$. (Note the 2 included here with the middle coefficient, as was usual in the 19th century.) Written as a product of matrices, this is

$$ax^2 + 2bxy + cy^2 = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \quad (11.2)$$

Clearly, $f(x, y)$ is symmetric under $(x, y) \rightarrow (-x, -y)$. Also, since x and y have equivalent status, (c, b, a) will achieve the same values as (a, b, c) .

Gauss made a major analysis of these forms as recorded in Section V of his *Disquisitiones Arithmeticae*. In this he distinguishes various types of quadratic form depending on whether the values it can assume for integers x and y , not both zero, are all strictly positive, strictly negative, or both positive and negative depending on (x, y) . There is a close parallel with the possible values of the quadratic $at^2 + 2bt + c$ in the one variable t depending on the discriminant $b^2 - ac$; compare with

$$a \left(\frac{x}{y} \right)^2 + 2b \left(\frac{x}{y} \right) + c = 0 \quad \text{with roots} \quad \frac{x}{y} = \frac{1}{a} (-b \pm \sqrt{b^2 - ac}).$$

In §171 of his monumental work Gauss showed that any positive definite binary quadratic form with real coefficients (that is, one whose value is greater than zero for all non-zero x and y) could be reduced to a canonical form $\mathcal{F} = AX^2 + 2BXY + CY^2 = (A, B, C)$ in which $|2B| < A \leq C$. In this reduced form, for integers x, y not both zero, the least value of $\mathcal{F}(x, y)$ is $\mathcal{F}(1, 0) = A$. The transformation to Gauss' reduced form is by a change of variable from x, y to X, Y using a matrix multiplication with determinant $+1$. Two forms connected by such a transformation are said to be 'properly equivalent'.

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} u & u' \\ v & v' \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}, \quad uv' - vu' = +1. \quad (11.3)$$

Since f is positive, the discriminant of the quadratic form is $b^2 - ac < 0$. Hence the determinant of the representation matrix $D = ac - b^2 > 0$. This value is unchanged by the change-of-axes transformation Eq 11.3¹². Here is a brief description of Gauss's reduction algorithm.

Step 1 : If $a > c$, interchange the values of a and c . This must be done by a proper equivalence transformation; the correct matrix multiplication is

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} c & -b \\ -b & a \end{pmatrix},$$

giving $(a' = c, -b, a)$. Note the transposed matrix on the left, giving the form $U^T M U$.

Step 2 : Replace the $-b$ by $b' = -b \bmod c > 0$ or by $b' = (-b \bmod c) - c < 0$, whichever has the smaller absolute value. This is equivalent to multiplication by another matrix pair $V^T M' V$:

$$\begin{pmatrix} 1 & 0 \\ n & 1 \end{pmatrix} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix} \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} c & -b + nc \\ -b + nc & a - 2nb + n^2c \end{pmatrix},$$

where the integer n is chosen so that $|b'| = |-b + nc|$ is as small as possible. The third coefficient $c' = a - 2nb + n^2c$ is equal to $(D + b'^2)/a'$. The algorithm gives (a', b', c') as a partial reduction of the form f . If this does not have the required properties that $|2b'| < a' \leq c'$, repeat the process until a derived form does. A positive definite binary quadratic form has a unique reduced form.

Example This example is given by Gauss, *D.A. §177*.

The initial given form is $304x^2 + 434xy + 155y^2$, or $(304, 217, 155)$, which has determinant $D = 304 \cdot 155 - 217^2 = 31$, meaning that the discriminant $-D = -31$. Because $304 > 155$, swap them over by Step 1. Next $-217 \bmod 155 = 93 \equiv -62$, so replace b by -62 since it has lower absolute value than 93. Finally replace the third coefficient by $(31 + (-62)^2)/155 = 3875/155 = 25$. (This will appear automatically if the matrix multiplication of Step 2 is applied.) The partially reduced form is $(155, -62, 25)$. A second cycle of reduction gives -62 replaced by $+62 \bmod 25 = 12$, and 25 replaced by $(31 + 12^2)/25 = 7$, giving $(25, 12, 7)$. Two more cycles are required to arrive at the required canonical form : $(7, 2, 5)$, then $(5, -2, 7)$. Note that $|2 \times (-2)| < 5 < 7$.

The minimum of the original form $304x^2 + 434xy + 155y^2$ is 5 at $(x, y) = (5, -7)$ whilst the minimum of the fully reduced form is also 5 at $(X, Y) = (1, 0)$. This is consistent with the transformation matrix

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 5 & -2 \\ -7 & 3 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \quad \text{with inverse} \quad \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 7 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

¹²The coefficients resulting from the transformation are $A = au^2 + 2buw + cv^2$, $B = auu' + b(uw' + u'v) + cvv'$, $C = au'^2 + 2bu'v' + cv'^2$. The transformation matrix represents an element of the modular group. Recall §7.

Essentially, the matrix changes the basis vectors used to navigate from the origin $(0, 0)$ to points at which the form is evaluated, but it does not change the values of the form at those points.

The Gauss reduced proper equivalent form (A, B, C) has $|2B| < A$ and $|2B| < C$. Therefore $4B^2 < AC$. Since the transformation to canonical form does not change D , we have $D = AC - B^2 > AC - AC/4 = \frac{3}{4}AC$. Moreover, $A \leq C$ so $D > \frac{3}{4}A^2$.

Building on Gauss's work, in 1826 Charles Hermite published an interesting consequence of this inequality which gives rise to a bound on the error in approximating any real number θ by a fraction u/v . He posed the carefully chosen form

$$f(x, y) = (x - \theta y)^2 + \frac{y^2}{\delta^2}$$

where δ is an arbitrary real. The discriminant is $-1/\delta^2$, which is negative, $-D$, say. Applying a transformation matrix as in Eq 11.3 reduces this to Gauss canonical form \mathcal{F} with coefficients and discriminant:

$$\begin{aligned} A &= (u - \theta v)^2 + \frac{v^2}{\delta^2}, & C &= (u' - \theta v')^2 + \frac{v'^2}{\delta^2}, \\ B &= (u - \theta v)(u' - \theta v') + \frac{vv'}{\delta^2}, & D &= \frac{(uv' - u'v)^2}{\delta^2} = \frac{1}{\delta^2}. \end{aligned}$$

Now take the inequality in §1.7, namely that for any real numbers r and s , $r^2 + s^2 \geq 2rs$ with equality only when $r = s$. Apply this to the two terms which make up coefficient A :

$$A = (u - \theta v)^2 + \frac{v^2}{\delta^2} > 2(u - \theta v)\frac{v}{\delta},$$

$$\text{so} \quad \left| 2(u - \theta v)\frac{v}{\delta} \right| < A < \sqrt{\frac{4D}{3}} = \sqrt{\frac{4}{3\delta^2}}.$$

$$\text{Finally} \quad |u - \theta v| < \frac{1}{\sqrt{3}v} \quad \text{implying} \quad \left| \frac{u}{v} - \theta \right| < \frac{1}{\sqrt{3}v^2} \quad (11.4)$$

which is the main result of this section. It means that *any* real number θ can be approximated by at least one fraction u/v to a relative accuracy greater than $1/(\sqrt{3}v^2)$. Since the convergents of θ are the best approximations, this can be further interpreted to mean that many convergents of θ should satisfy Eq 11.1 with $\mathcal{A} = \sqrt{3}$ and $\alpha = 2$. Indeed, as will be described in §11.4, the numerical evidence is that over 80% of convergents have $\epsilon_k q_k^2 < 1/\sqrt{3}$.

11.3 Hurwitz's three-convergents theorem

We continue to look for tighter bounds on the coefficient \mathcal{A} in Eq 11.1. Item 2 of the list in §11.1 (Eq 1.11) was derived by considering the bounds on errors of two adjacent convergents, and concluded that at least one in two convergents satisfy $\epsilon_k q_k^2 < 1/\mathcal{A}$ for $\mathcal{A} = 2 = \sqrt{4}$. It seems natural to extend this to asking what constraints might apply to three or more consecutive convergents. Joseph Lagrange and Adolf Hurwitz were amongst the first to study this. In 1891 Hurwitz showed that, in any three consecutive convergents to $\theta \in \mathbb{R}$, there is at least one with $\epsilon_k < 1/(\sqrt{5}q_k^2)$. Moreover, because all *irrationals* have an infinity of convergents, Hurwitz added this corollary: for any irrational θ there is an infinity of rationals p/q (at least one convergent in three) such that

$$|\epsilon| = \left| \frac{p}{q} - \theta \right| \leq \frac{1}{\sqrt{5}q^2}. \quad (11.5)$$

Moreover, $\sqrt{5}$ is the largest number for which this statement is true.

In their classic book Hardy and Wright prove Hurwitz's theorem, and conclude from it that the most 'difficult' irrational number to approximate by a rational is the Golden Ratio, $G = \{1 : \underline{1}\} = \frac{1}{2}(1 + \sqrt{5})$, because it converges at the slowest rate of all continued fractions. We stated this intuitively in §1.1. The equality in the \leq sign of Eq 11.5 occurs for G and numbers equivalent to G . If $\mathcal{A} > \sqrt{5}$ (meaning a tighter bound on error), there are some $\theta \in \mathbb{R}$, including G , for which there is no rational p/q , or are only a finite number, with error less than $1/(\mathcal{A}q^2)$.

Here is my own proof of Hurwitz's theorem, which draws on the fundamental relations in §1.5 and the terminology of Figure 2, Part I. Consider the question 'for what values of \mathcal{A} are the errors in three consecutive convergents all greater than $1/(\mathcal{A}q_k^2)$?'.

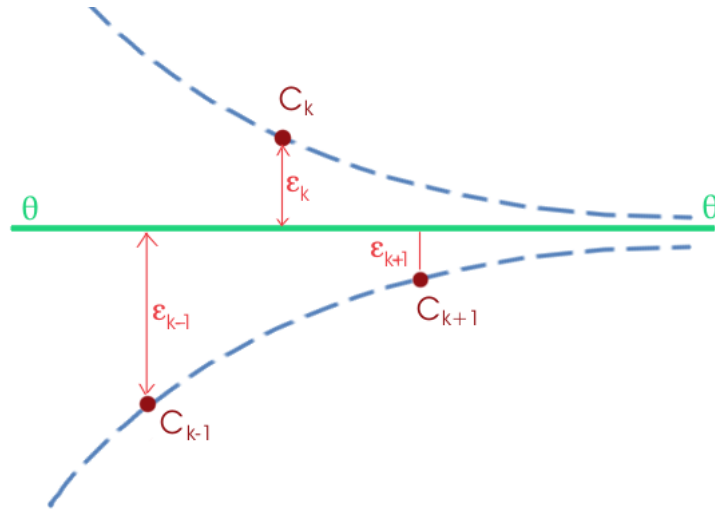


Figure 18: Three adjacent convergents as a function of k

Look at Figure 18. If all three errors, ϵ_{k-1} , ϵ_k , ϵ_{k+1} , are each to have the largest possible value, the two dashed curves will need to be almost horizontal lines. Strictly horizontal is not possible since $q_{k-1} < q_k < q_{k+1}$. The closest that the curves joining the odd and even convergents can be to horizontal is when all the partial quotients a_k are as small as they can be – that is, they are all 1. That corresponds to the continued fraction $\{1, \underline{1}\} = \frac{1}{2}(1 + \sqrt{5})$. This picture gives rise to the following line of proof.

We are investigating the cases where

$$\epsilon_{k-1} \geq \frac{1}{\mathcal{A}q_{k-1}^2}, \quad \epsilon_k \geq \frac{1}{\mathcal{A}q_k^2}, \quad \text{and} \quad \epsilon_{k+1} \geq \frac{1}{\mathcal{A}q_{k+1}^2}. \quad (11.6)$$

From Eq 1.6

$$\Delta_k \equiv C_k - C_{k-1} = |\epsilon_k| + |\epsilon_{k-1}| = \frac{1}{q_{k-1}q_k} \geq \frac{1}{\mathcal{A}q_{k-1}^2} + \frac{1}{\mathcal{A}q_k^2}, \quad (11.7a)$$

$$\Delta_{k+1} \equiv C_{k+1} - C_k = |\epsilon_{k+1}| + |\epsilon_k| = \frac{1}{q_k q_{k+1}} \geq \frac{1}{\mathcal{A}q_k^2} + \frac{1}{\mathcal{A}q_{k+1}^2}. \quad (11.7b)$$

Rearrange Eq 11.7 a, b and take the borderline situation of equality

$$q_k^2 - (\mathcal{A}q_{k-1})q_k + q_{k-1}^2 = 0, \quad q_k^2 - (\mathcal{A}q_{k+1})q_k + q_{k+1}^2 = 0.$$

These quadratics each have two solutions, but $q_{k-1} < q_k < q_{k+1}$ so the correct choice of signs is

$$\frac{q_k}{q_{k-1}} = \frac{1}{2}(\mathcal{A} + \sqrt{\mathcal{A}^2 - 4}) > 1, \quad \frac{q_k}{q_{k+1}} = \frac{1}{2}(\mathcal{A} - \sqrt{\mathcal{A}^2 - 4}) < 1.$$

Note that $\mathcal{A} \geq 2$ for a real root; compare this with the $1/(2q_k^2)$ criterion for a convergent. The reciprocal of the last (right) equation is

$$\frac{q_{k+1}}{q_k} = \frac{1}{2}(\mathcal{A} + \sqrt{\mathcal{A}^2 - 4}) > 1.$$

The recursion relation states that $q_{k+1} = a_{k+1}q_k + q_{k-1} \geq q_k + q_{k-1}$ because $a_{k+1} \geq 1$. Therefore

$$\begin{aligned} \frac{q_{k+1}}{q_k} &\geq 1 + \frac{q_{k-1}}{q_k} \\ \frac{1}{2}(\mathcal{A} + \sqrt{\mathcal{A}^2 - 4}) &\geq 1 + \frac{1}{2}(\mathcal{A} - \sqrt{\mathcal{A}^2 - 4}) \end{aligned} \quad (11.8)$$

The solution is $\mathcal{A} \geq \sqrt{5}$. This is the condition for any three consecutive convergents to any θ to have large errors, as in Eq 11.6. The logical converse statement is this: the condition for at least one in three convergents to have $\epsilon < 1/(\mathcal{A}q^2)$ is that $\mathcal{A} < \sqrt{5}$. If θ is irrational, this essentially proves Hurwitz's theorem.

The borderline case $\mathcal{A} = \sqrt{5}$ corresponds to $\theta = G = \{1, \underline{1}\}$. If you refer back to Table 21 in §10.1, you will see that $1/\sqrt{5}$ characterises the asymptotic convergence of error for G . §8.3 explains equivalent numbers, stating that they have the same tail to their a_k sequences of partial quotients. Therefore, if $\mathcal{A} > \sqrt{5}$, no number equivalent to G will have one in three convergents meeting $\epsilon < 1/(\mathcal{A}q^2)$. Perhaps a finite number of rational approximations may achieve this limit on error, but certainly not an infinity.

Suppose that we exclude numbers which are equivalent to G . Does that allow a reduction (a tightening) of the bound on $\epsilon_k q_k^2$ from $1/\sqrt{5}$? This is a deep question, which leads to a discussion of the Lagrange-Markoff Spectrum. Since this is such a large topic, it has a section to itself – Section 12. Before getting into that, however, the next sub-section describes some results of my numerical investigation of the percentage of convergents lying below a variety of specified limits. This generally supports the analytic evidence presented above.

11.4 Numerical evidence for error estimates

This subsection returns to the estimates of error ϵ_k of the k^{th} convergent of a generic continued fraction, presented in §1.7, Part 1, and describes numerical evidence to support the analysis. Errors for some special cases of quadratic irrationals, with recurring sequences of a_k , were examined in §10. Though Tables 1 and 2 give some numerical support for the algebraic inequalities of §1.7, this present section reports more thorough numerical investigation. Specific statements from §1.7 tested here include:

1. An upper bound on error is estimated by

$$|\Delta_{k+1}| = \frac{1}{q_{k+1}q_k} = \frac{1}{(a_{k+1}q_k + q_{k-1})q_k} < \frac{1}{a_{k+1}q_k^2} \leq \frac{1}{q_k^2}. \quad \text{Copy of (1.9)}$$

$$\text{so } \epsilon_k < \frac{1}{q_k^2}. \quad \text{Copy of (1.10)}$$

2.

$$\epsilon_k + \epsilon_{k+1} < \frac{1}{2q_k^2} + \frac{1}{2q_{k+1}^2} \quad \text{Copy of (1.11a)}$$

$$\text{so if } \epsilon_{k+1} > \frac{1}{2q_{k+1}^2}, \text{ then } \epsilon_k < \frac{1}{2q_k^2} \text{ and vice versa.} \quad \text{Copy of (1.11b)}$$

Hence no two adjacent convergents can both have errors greater than $1/(2 \times \text{denominator}^2)$.

3. Lower and upper bounds in terms of the next partial quotient a_{k+1} are

$$\frac{1}{(a_{k+1} + 2)q_k^2} < \epsilon_k < \frac{1}{a_{k+1}q_k^2}. \quad \text{Copy of (1.14)}$$

Bear in mind that expressions for the *exact* error in a convergent are Eq 1.8, §1.7:

$$\epsilon_k = \frac{\rho_k}{q_k(q_k + \rho_k q_{k-1})}. \quad \text{Copy of (1.8)}$$

which can also be stated as Eq. 2.3, §2.4:

$$\epsilon_k q_k^2 = \frac{1}{\theta_{k+1} + \chi_k} = \frac{1}{\{a_{k+1} : a_{k+2}, \dots, \} + \{0 : a_k, a_{k-1}, \dots, a_2, a_1\}} \quad \text{Copy of (2.3)}$$

All the above bounds and other inequalities are derived from these.

I have carried out two separate numerical experiments to assess the error in convergents relative to $1/q_k^2$ and test the above analytical inequalities. In the first I compiled a table of 2000 convergents calculated from 141 random fractions. Each fraction p/q was produced by creating a random integer with 7, 8 or 9 digits for p and another such for q . The continued fraction expansion of each was calculated and the full sequence of convergents evaluated. The first, integer convergent C_0 was discarded, as were the highest one or two, C_F and C_{F-1} because of poor machine accuracy in calculating error. So typically convergents from C_1 to C_{14} were recorded for 141 fractions, giving 2000 convergents in all. Table 24 lists the results against various criteria. The columns record the number and percentage of the 2000 convergents for which $\epsilon_k < f/q_k^2$ where f is the fraction in the column label. Thus, column 2 records that 1430, being 71.4% of convergents, have error less than $\frac{1}{2q_k^2}$. The first four columns are points on the cumulative distribution of errors relative to $\frac{1}{2q_k^2}$.

Table 24: First survey of number and % of convergents with errors $\epsilon_k q_k^2 < f$.

$f = \frac{1}{A}$	< 1	$< \frac{1}{2}$	$< \frac{1}{3}$	$< \frac{1}{4}$	$< \frac{1}{a_{k+1}}$	$< \frac{1}{a_{k+1} + 2}$
Number	2000	1430	946	686	2000	0
Percentage %	100	71.4	47.3	34.3	100	0

Note these points which together show that this first study fully supports all formulae and inequalities in §1.7:

1. Column 1 shows that 100% of convergents satisfy Eq 1.10.

2. In column 2 the 71.4% figure is to be compared with the theorem in Eq 1.11 that at least 50% must have such a low error; so in fact the theorem is conservative in its estimate.
3. With respect to column 2, I also searched for any instances where two adjacent convergents both had errors greater than $1/(2 \times \text{denominator}^2)$, but there was none.
4. In the last two columns, all convergents have error less than $\frac{1}{a_{k+1}q_k^2}$, and none has error less than $\frac{1}{(a_{k+1}+2)q_k^2}$, consistent with Eq 1.14.

The second study was made at a later date using arbitrary precision software. In all about 650 random real numbers were generated, each in the range 0 to 1, and with 39 to 42 digits. The first 35 convergents were calculated and C_{34} taken to be the exact value θ . Statistics were collected from C_0 to C_{30} inclusive, making a data set of over 20,000 convergents. The results are presented in Table 25 and in Figure 19.

Table 25: Second survey of number N and percentage of convergents with errors $\epsilon_k q_k^2 < f$.

f	< 1	$< \frac{9}{10}$	$< \frac{3}{4}$	$< \frac{1}{\sqrt{3}}$	$< \frac{1}{2}$	$< \frac{1}{\sqrt{5}}$	$< \frac{1}{\sqrt{8}}$	$< \frac{1}{3}$	$< \frac{1}{4}$	$< \frac{1}{5}$	$< \frac{1}{6}$	$< \frac{1}{10}$
N	20057	19840	18864	16285	14369	12807	10105	9535	7104	5676	4759	2872
$\%$	100	98.9	94.1	81.2	71.6	63.9	50.4	47.5	35.4	28.3	23.7	14.3

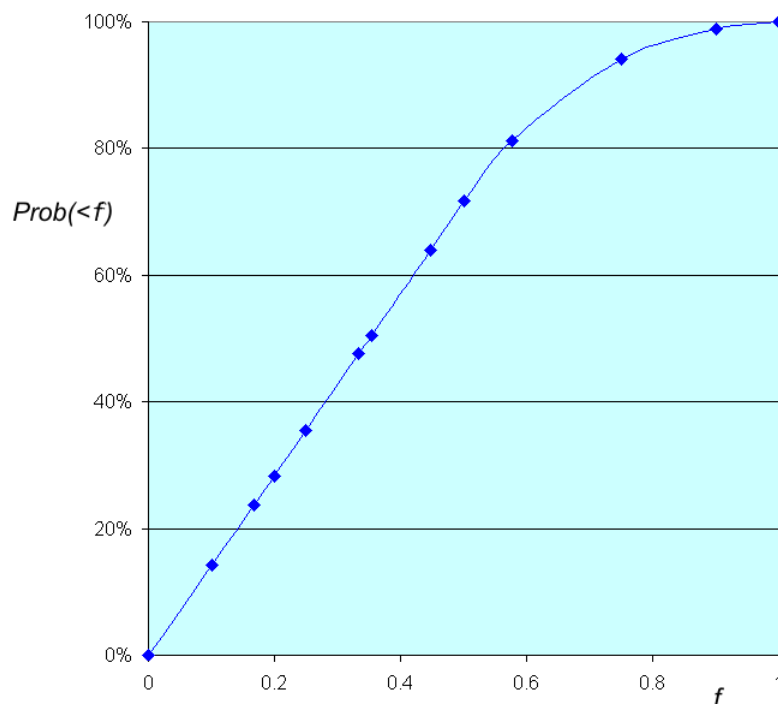


Figure 19: Cumulative probability of $\epsilon_k q_k^2 < f$ for a set of over 20,000 convergents.

The graph, which is the cumulative probability that $\epsilon_k q_k^2 < f = 1/\mathcal{A}$ as a function of f , is remarkable straight up to $f = 0.6$, and over this straight section the gradient is 1.42 , almost $\sqrt{2}$. Indeed, $\sqrt{(2/3)} = 0.8165$, to be compared with the 81.2% in the $1/\sqrt{3}$ column of Table 29. Over the

straight section the probability that the relative error $\epsilon_k q_k^2$ will be between f_0 and $f_0 + \delta f$ is almost $\delta f \sqrt{2}$ and independent on f_0 .

Rearranging Eq 1.8

$$\epsilon_k q_k^2 = \frac{\rho_k}{1 + \rho_k \frac{q_{k-1}}{q_k}} \approx \rho_k, \quad (11.9)$$

so the cumulative probability of obtaining obtaining $\epsilon_k q_k^2$ should be about equal to the probability of obtaining a value less than ρ as the k^{th} remainder. We will see in Part IV that cumulative probability $\mathcal{P}(\rho)$ is actually a logarithmic function, but it does not depart strongly from a straight line, corresponding to a constant frequency (uniform distribution) for ρ . This goes some way towards explaining the substantially linear relationship in Figure 19.

Since $\rho_k = 1/\theta_{k+1} = 1/(a_{k+1} + \rho_{k+1})$, Eq 1.8 can also be written

$$a_{k+1} \epsilon_k q_k^2 = \frac{a_{k+1}}{(a_{k+1} + \rho_{k+1})} \frac{1}{\left(1 + \rho_k \frac{q_{k-1}}{q_k}\right)} \approx \frac{1}{1 + \rho_k \frac{q_{k-1}}{q_k}} \approx \text{constant}, H < 1. \quad (11.10)$$

If we refer back to §10.1.2, dealing with the simplest continued fraction $\{a, \underline{a}\}$, formula Eq 10.7.b quantifies the unsigned error as exactly as constant with value

$$\epsilon_k = \frac{1}{\sqrt{a^2 + 4}} \frac{1}{q_k^2}.$$

I have looked for evidence of a constant H in the 2000 values of convergents which make up Table 24 (first study), and later more thoroughly in the second, arbitrary precision study. Broadening the search for simple models, Eq 1.8 can also suggest that

$$\epsilon_k q_k^2 = \frac{1}{(a_{k+1} + \rho_{k+1})} \frac{1}{\left(1 + \rho_k \frac{q_{k-1}}{q_k}\right)} \approx \frac{1}{a_{k+1} + h} \quad (11.12)$$

for $h < 1$ approximately constant. Accordingly I have investigated these two models :

Model 1 : Simple proportionality

$$\epsilon_k = \frac{H}{a_{k+1} q_k^2}$$

Model 2 : Linear denominator

$$\epsilon_k q_k^2 = \frac{1}{a_{k+1} + h}.$$

From the data in Study One I calculated average values of H and h from these formulae. H has a moderate spread about mean 0.713 , standard deviation 0.172 , with median 0.724 . I find the mean and median tantalisingly close to $1/\sqrt{2} = 0.707$ and ever closed to $1/(2 \ln 2) = 0.721$. In Model 2, h varies more about its mean: 0.903 ± 0.427 , with median also 0.903 .

Calculation of $a_{k+1} \epsilon_k q_k^2$ for the second data set of 20,057 convergents gave mean and standard deviation in Model 1 as $H \approx 0.72 \pm 0.18$. Extreme values from 0.46 to 0.98 were observed. I did not re-examine Model 2.

Interpolation on the first four columns of Table 24 gives that 50% of convergents have error below $0.352/q_k^2$. This tallies with Study 2, Table 25, where the median is at $1/\sqrt{8} = 0.354$. Equating this with the median of H gives a representative value of a_{k+1} to be $0.724/0.352 = 2.25$. Similarly,

from Model 2 we get a ‘median’ value for a_{k+1} to be $1 \cdot 94$. Moreover, the median value of partial quotient a_k in the data set is 2. As further evidence, the value $a = 2$ makes $a/\sqrt{a^2 + 4}$ (the formula for a single recurring partial quotient) equal to $1/\sqrt{2}$. So it seems that a ‘typical’ value for partial quotients is about 2.

To look for any dependency of H and h on a_k, a_{k+1} I have searched the data from Study One using the Eureka adaptive function-fitting package from Cornell University. This yielded three models for H with increasing complexity:

$$\text{i) } H = 0 \cdot 713, \quad \text{ii) } H = \frac{a_{k+1}}{a_{k+1} + 0 \cdot 85} \quad \text{iii) } H = \frac{a_{k+1}}{a_{k+1} + \frac{0 \cdot 774}{a_k} + 0 \cdot 408} .$$

ii) is equivalent to the linear denominator Model 2 above. Overall, I suggest that the most useful, practical estimates of actual error (as opposed to bounds on error) are

$$\text{Model 1) } \quad \epsilon_k \approx \frac{1}{2 \ln 2 \, a_{k+1} \, q_k^2} , \quad \text{Model 2) } \quad \epsilon_k \approx \frac{1}{(a_{k+1} + 0 \cdot 9) \, q_k^2} . \quad (11.13)$$

The two estimates are equal when $a_{k+1} \approx 2 \cdot 2$. Also $(a + 0 \cdot 9)$ and $\sqrt{a^2 + 4}$ agree for $a \approx 1.8$, again suggesting that a typical value of a_k is about 2. More will be said about typical values of the partial quotients in Part IV.

Of course, both these Models involve knowing the next partial quotient a_{k+1} since the only way of estimating error is to compare with the next level of approximation. But once you have calculated a_{k+1} , you might as well calculate C_{k+1} and have a better approximation to θ than the convergent C_k whose error this Section has been estimating!

12 The Lagrange-Markoff spectrum

We pick up from §11.3 on Hurwitz's three-convergents theorem which states that $1/\sqrt{5}$ is a critical value of the relative error $\epsilon_k q_k^2$, separating the reals into two classes according to whether they have a finite or an infinite number of rational approximations with errors $\epsilon_k q_k^2 \leq 1/\mathcal{A}$, $\mathcal{A} = \sqrt{5}$. Suppose that all numbers equivalent to $G = \frac{1}{2}(1 + \sqrt{5})$ are excluded from the reals. Does that allow a tightening of the bound on $\epsilon_k q_k^2$ for the remaining numbers?

The smallest perturbation from $\{1 : \underline{1}\}$ would be to introduce a sparse sprinkling of '2's into the tail of the a_k sequence. Since the tail of infinitely recurring '1' in G is to be perturbed to produce a non-equivalent number, an infinite number of '2's must be introduced, no matter how sparse. At each isolated 2 a trough down to $0.309 = 1/(2G)$ will occur in the value of $\epsilon_k q_k^2$ as shown in Figure 12. Therefore, every irrational whose tail of a_k has a sprinkling of isolated 2 values amongst a background of 1s will have an infinite number of convergents which satisfy $\epsilon_k q_k^2 = 1/(2G)$.

At the other extreme, if the recursion sequence were entirely 2s, the value of $\epsilon_k q_k^2$ would be $0.35355 = 1/\sqrt{8}$ for all convergents C_k . However, as in Figure 14, where the 2s are in blocks separated by at least one 1, there is a dip down to about 0.33. Indeed, examination of all the panels in Figures 14, 15 and 16 will show many cases of dips down to about 1/3. The point is that, in all numbers θ equivalent to those in the panels, there is will an infinite number of convergents which share this dip to about 1/3. Therefore the constant \mathcal{A} in Eq 11.1 could be set a limit near 1/3, whose precise value is yet to be determined, and still there be an infinity of fractions p_k/q_k with relative error $\epsilon_k q_k^2 < 1/\mathcal{A}$. It turns out, as we shall see shortly, that there is not a single limit $1/\mathcal{A}$ but rather a discrete sequence of limits $1/\mathcal{A}_M$, getting closer to each other, whose limit point is 1/3. Each limit, indexed by \mathcal{M} , in turn separates out a class of numbers for which an infinity of rational approximations exists to an accuracy of $1/\mathcal{A}_M$, but no better. This sequence is called the Lagrange Spectrum or Lagrange-Markoff Spectrum.

The spectrum was described first by Lagrange in about 1790, and analysed extensively over the next 150 years, especially by the Russian Andrei Markoff who published two influential papers of 1879 and 1880 (written in French and published in *Mathematische Annalen*). Markoff (also written Markov) was also a pioneer of theory of stochastic processes and gave his name to the 'Markov chain'. There is an extensive literature of the Lagrange-Markoff Spectrum because it is one of those topics in maths which can be looked at from many different points of view, and connected to areas of maths which otherwise seem disparate. I will indicate some of these at the end of the section. Until the 20th century the literature was concerned only with the Spectrum to the limit point $\mathcal{A} = 3$, up to which it has discrete values. Later studies have extended to $\mathcal{A} > 3$ where the behaviour is complicated, showing gaps such as between $\mathcal{A} = \sqrt{12}$ and $\sqrt{13}$, and also regions of chaotic behaviour. Here I will stick to the classical region where analysis of the Spectrum is essentially analysis of the transition from $\{1 : \underline{1}\}$ to $\{2 : \underline{2}\}$, initiated in §10.4. All numbers of concern can be represented by continued fractions in which the partial quotients are a mixture of '1's and '2's .

12.1 Numerical evidence for the Spectrum

I must ask the reader to accept for the time being that all the critical values of \mathcal{A}_M are associated with reals whose continued fractions have a recurring sequence of partial quotients – that is, which are quadratic surds. Later I will offer some justification for this. Accepting this, therefore, let us return to §10.4, Figures 14, 15, 16, and examine with high precision the values of $\epsilon_k q_k^2$ for continued fractions whose partial quotients are recurring sequences made only of 1s and 2s.

No. of 1s	1	2	3	4	5	6
No. of 2s						
1	0.288675	0.316228	0.306186	0.310087	0.308607	0.309173
2	0.325396	0.336336	0.332182	<i>0.333772</i>	0.333165	<i>0.333397</i>
3	0.319505	0.332820	0.327781	0.329713	0.328976	0.329257
4	0.320527	<i>0.333421</i>	0.328538	0.330409	0.329695	0.329968
5	0.320352	0.333318	0.328408	0.330290	0.329572	0.329846
6	0.320382	<i>0.333336</i>	0.328430	0.330310	0.329593	0.329867
7	0.320376	0.333333	0.328427	0.330307	0.329589	0.329864
8	0.320377	<i>0.333333</i>	0.328427	0.330307	0.329590	0.329864

Table 26: Minimum (absolute) values of ϵq^2 for fractions with recurring partial quotients sequences of the form $\{0 : \underline{1_s 2_t}\}$.

I have computed the values of $\epsilon_k q_k^2$ for about 40 convergents, as for Figures 14, 15, 16, and determined the minimum value for continued fractions from $\{0 : \underline{1_1 2_1}\}$ to $\{0 : \underline{1_{14} 2_{14}}\}$. Table 26 reports the most important part of the results, and the data are also plotted in Figure 20. Note the undulation in the graphs, with sequences having the 1s and 2s in multiples of 2 having greater values than where there are odd numbers of 1 or 2. The largest number in Table 26 is 0.336336 , corresponding to $\{1, 1, 2, 2\} = (-1 + \sqrt{(221)})/10$.

If a ceiling bound value is to be set so that there are an infinity of convergents lying under this ceiling, the ceiling must be high enough. For this reason, its value must be the largest number in Table 26. The logic develops as follows:

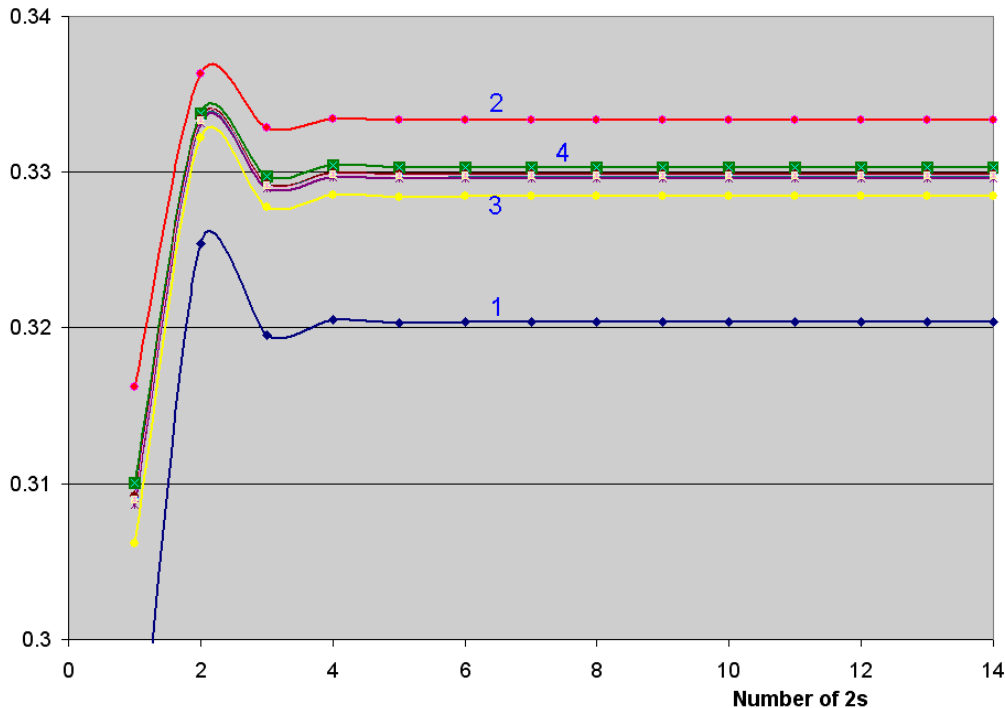


Figure 20: Plots of data in Table 26 showing minimum values of fractions of the form $\{1_m 2_n\}$. Numbers on lines state number of 1s.

1. from Hurwitz' theorem, every real number has an infinity of rational approximations $u/v = p_k/q_k$ which satisfy $\epsilon q^2 < 1/\sqrt{5}$. The limiting θ are equivalent to G ,
2. if all reals equivalent to G are excluded, the ceiling on error can be lowered to $1/\sqrt{8}$, at which every remaining real number has an infinity of rational approximations $u/v = p_k/q_k$ which satisfy $\epsilon q^2 < 1/\sqrt{8}$. The limiting θ now are equivalent to $-1 + \sqrt{2}$, the so-called Silver Mean.
3. if all reals equivalent to either $\{0 : \underline{1}\}$ or $\{0 : \underline{2}\}$ are excluded, the ceiling bound on error can be reduced yet again to 0.336336 . The limiting θ now are equivalent to $\{0 : \underline{1, 1, 2, 2}\}$,
4. the process of successively excluding equivalence classes corresponding to the next largest number in Table 26 can be continued indefinitely.

The next highest minima are picked out in italics in Table 26, and tabulated in Table 27. The sequence of decimal values in Table 27 is known as the 'Lagrange-Markoff Spectrum'. As originally defined, it applies only to continued fractions whose partial quotients are 1 or 2. The limiting value of the spectrum is exactly $1/3$. The spectrum is like an infinite sequences of numerical sieves of increasing fineness, the coarsest having a mesh size of $1/\sqrt{5}$. In stages these sieve out, from the numbers remaining, those which are the most difficult to approximate by rationals. After the first two or three stages of filtering there is so little difference in the mesh size (it is very close to $1/3$) that the spectrum has no more practical value as a guide to approximation, but it remains a mathematical curiosity and has been much studied as such.

We can also ask, as a corollary, whether there are any numbers whose partial quotients are entirely 1s and 2s which can pass through all stages of sieving and still have an infinity of rational approximations satisfying $\epsilon_k q_k^2 < 1/3$? The answer must be 'Yes' because this section opened by stating that fractions with a sufficiently sparse distribution of isolated 2 partial quotients, the vast majority being 1, will have a dip to $\epsilon_k q_k^2 = 1/(2G) = 1/(1 + \sqrt{5})$ at every 2 (Figure 12). Moreover, in a numerical experiment I have tested many numbers whose a_k are randomly 1 and 2 in nominally equal proportions, and found *many* convergents satisfying $\epsilon_k q_k^2 < 1/3$.

Our immediate task is to determine analytic expressions for the Spectrum values listed in Table 27. These numbers must be related to the θ . Referring back to Table 21 in §10, this shows that the coefficient \mathcal{A} (bottom line) is a pure square root – essentially the same square root as in θ (top line). So we can try multiplying the decimals in Table 27 by the square root featuring in the corresponding θ . For example, in row 3, $0.3363361485 \times \sqrt{221} = 4.9999963 \approx 5$, and in row 4 $0.3337723632 \times \sqrt{1517} = 12.9999944 \approx 13$. Very promising!

Here's another clue. 221 is close to $225 = 15^2$, and in the proof of Hurwitz' theorem, Eq 11.8, a square root with the form $\sqrt{D} = \sqrt{m^2 - 4}$, $m \in \mathbb{N}$ featured. Now $221 = 3^2 5^2 - 4$. Moreover, $1517 = 3^2 13^2 - 4$, so the same numbers 5 and 13 are appearing a second time. Following this lead takes us to Table 28. This corresponds to Table 27, with the minima listed in decreasing order. The analytic expression for each decimal is M/\sqrt{D} . The last three columns show in stages how the integers $M = 1, 2, 5, 13, \text{etc.}$ also consistently appear in the square root¹³. You may think it remarkable that the 'filter' numbers of the Lagrange-Markoff spectrum all involve periodic continued fractions which correspond to quadratic irrationals, but that is the case.

¹³Some 'adjustment' has been made in rows numbered 2 and 6 to correct for the cancellation between numerator and denominator which occurred in evaluating the continued fraction. This adjustment has been to multiply the square root by an integer. As a check that this is correct, observe that in row 2, $2/\sqrt{32} = 1/\sqrt{8}$.

	sequence	value	θ
1	$1_\infty 2_0$	$0.4472 = \frac{1}{\sqrt{5}}$	$\frac{1}{2}(-1 + \sqrt{5})$
2	$1_0 2_\infty$	$0.3536 = \frac{1}{\sqrt{8}}$	$(-1 + \sqrt{2})$
3	$1_2 2_2$	0.336336148521	$\frac{1}{10}(-9 + \sqrt{221})$
4	$1_4 2_2$	0.333772363220	$\frac{1}{26}(-23 + \sqrt{1517})$
5	$1_2 2_4$	0.333421446118	$\frac{1}{58}(-53 + \sqrt{7565})$
6	$1_6 2_2$	0.333397348289	$\frac{1}{17}(-15 + 5\sqrt{26})$
7	$1_8 2_2$	0.333342657413	$\frac{1}{178}(-157 + \sqrt{71285})$
8	$1_2 2_6$	0.333335890841	$\frac{1}{338}(-309 + \sqrt{257045})$
9	$1_{10} 2_2$	0.333334697180	$\frac{1}{466}(-411 + \sqrt{488597})$
10	$1_2 2_8$	0.333333371507	$\frac{1}{1970}(-1801 + \sqrt{8732021})$
11	$1_2 2_{10}$	0.333333335444	$\frac{1}{11482}(-10497 + 5\sqrt{11865269})$
12	$1_2 2_{12}$	0.333333333395	$\frac{1}{66922}(-61181 + \sqrt{10076746685})$
13	$1_2 2_{14}$	0.333333333335	$\frac{1}{390050}(-356589 + \sqrt{342312755621})$

Table 27: Maximum values in set of minima in Table 26 listed in decreasing order, with recursion sequence of partial quotients and θ .

The conclusion is that the coefficient $1/\mathcal{A}$ in Eq 1.11 can be set in turn to the values

$$\frac{1}{\sqrt{5}}, \quad \frac{2}{\sqrt{32}}, \quad \frac{5}{\sqrt{221}}, \quad \frac{13}{\sqrt{1517}}, \quad \frac{29}{\sqrt{7565}} \quad \dots$$

with general form $\frac{M}{\sqrt{9M^2 - 4}}$, $M = 1, 2, 5, 13, 29, 34, 89, 169, 233, \dots$ (12.1)

There are almost certainly values missing beyond this point in Tables 27 and 28¹⁴. The analysis has come to the point where we ask ‘What is this sequence of numbers, and where does it come from?’

12.2 Continued fractions and the Spectrum

To be clear, in the Lagrange-Markoff Spectrum we are looking to separate classes of reals into those which have an infinite number of rational approximations satisfying $\epsilon_k q_k^2 < 1/\mathcal{A}_M$ from those which only have a finite number, or even no such approximation. We are therefore seeking to identify: a) for a given θ , the minimum values of $\epsilon_k q_k^2$ as k ranges over all convergents, and b) those reals, θ , which have the greatest minima. Then, if the bound $1/\mathcal{A}_M$ is set to this greatest minimum, all θ with lower minima will also have an infinity of rational approximations.

The evidence of §12.1 is that there is a sequence of limits $1/\mathcal{A}_M$, but does not explain sufficiently why the \mathcal{A}_M correspond to particular patterns of partial quotients. Some light is shed by using the expression Eq 2.3 of §2.4 for the relative error:

$$\epsilon_k q_k^2 = \frac{1}{\theta_{k+1} + \chi_k},$$

$$\theta_{k+1} = \{a_{k+1} : a_{k+2}, \dots\}, \quad \chi_k = \{0 : a_k, a_{k-1}, \dots, a_2, a_1\}. \quad \text{Copy of (2.3)}$$

¹⁴There is also one intermediate between rows 8 and 9 in Table 28, but it has a different pattern of 11 and 22 pairs: $\{0 : \underline{2211221}_4\}$ has $M = 194$, $D = 338720$.

	M	D	$D + 4$	$(D + 4)/3^2$	$\sqrt{(D + 4)}/3$
1	1	5	9	1	1
2	2	32	36	4	2
3	5	221	225	25	5
4	13	1517	1521	169	13
5	29	7565	7569	841	29
6	34	10400	10404	1156	34
7	89	71285	71289	7921	89
8	169	257045	257049	28561	169
9	233	488597	488601	54289	233
10	985	8732021	8732025	970225	985
11	5741	296631725	296631729	32959081	5741
12	33461	10076746685	10076746689	1119638521	33461
13	195025	342312755621	342312755625	38034750625	195025

Table 28: Expressions for the Markoff ratios, and identification of Markoff numbers in the Spectrum. D is the integer inside the $\sqrt{\quad}$ in Table 27.

Bear in mind that the tail of the continued fraction for θ contains an infinite number of 1s and 2s. We examine the effect that various groupings of 1s and 2s would have on the maximum value of $\theta_{k+1} + \chi_k$, and hence on the minimum of $\epsilon_k q_k^2$. Though χ_k is a finite fraction, we will follow Markoff and assume that it is has sufficient partial quotients to be treated as if infinite. This corresponds to the split being made in the a_k sequence of θ at a high value of k , in the tail of the continued fraction.

As typical examples, take the following three groupings within an infinite sequence of partial quotients. Each grouping has two or three 2s in an otherwise long chain of 1s extending either side of the 2s. To clarify the notation I will drop the commas from between the partial quotients.

Grouping 1: $\theta = \{ \text{containing } \dots 1_m 2 2 1_n \dots \}$, m, n arbitrary positive integers. The grouping can be cut at five significant places to form $\mathcal{A} = \theta_{k+1} + \chi_k$. The cut positions are $\{ \dots 1_A 1_B 2_C 2_D 1_E 1 \dots \}$

$$\begin{aligned}
\text{A :} & \quad \{1: 2 2 1, \dots\} + \{0: 1 1 1 1, \dots\} \\
\text{B :} & \quad \{2: 2 1 1, \dots\} + \{0: 1 1 1 1, \dots\} \\
\text{C :} & \quad \{2: 1 1 1, \dots\} + \{0: 2 1 1 1, \dots\} \\
\text{D :} & \quad \{1: 1 1 1, \dots\} + \{0: 2 2 1 1, \dots\} \\
\text{E :} & \quad \{1: 1 1 1, \dots\} + \{0: 1 2 2 1, \dots\}
\end{aligned}$$

Recall from §1.3 that, for a given continued fraction, the value of θ is increased if $a_k = 1$ is replaced by $a_k = 2$ wherever k is even; conversely, the value of θ is decreased if $a_k = 1$ is replaced by $a_k = 2$ wherever k is odd. Using this the values of \mathcal{A} in A to E above can be put in order by inspection, without evaluating any of them. For instance, B and C have the same fractional part – only their integer parts are swapped over. Therefore

$$B = C > E > A = D.$$

Compare these relations to the graphs in Figure 17. The maximum is at convergents B and C, corresponding to a minimum in $\epsilon_k q_k^2$.

Grouping 2: $\theta = \{\dots 1_m 2 1 2 1_n \dots\}$. The sequence a_k can be cut at six significant positions $\{\dots 1_A 1_B 2_C 1_D 2_E 1_F 1 \dots\}$

$$A : \quad \{1:2 1 2 1 \dots\} + \{0:1 1 1 1 1 \dots\}$$

$$B : \quad \{2:1 2 1 1 \dots\} + \{0:1 1 1 1 1 \dots\}$$

$$C : \quad \{1:2 1 1 \dots\} + \{0:2 1 1 1 1 \dots\}$$

$$D : \quad \{2:1 1 1 \dots\} + \{0:1 2 1 1 1 \dots\}$$

$$E : \quad \{1:1 1 1 \dots\} + \{0:2 1 2 1 1 \dots\}$$

$$F : \quad \{1:1 1 1 \dots\} + \{0:1 2 1 2 1 \dots\}$$

$$B = D > F > A = E > C.$$

The maximum is at convergents B and D.

Grouping 3: $\theta = \{\dots 1_m 2 2 2 1_n \dots\}$. The sequence can be cut at the seven significant positions $\{\dots 1_A 1_B 1_C 2_D 2_E 2_F 1_G 1 \dots\}$

$$A : \quad \{1:1 2 2 2 1 \dots\} + \{0:1 1 1 1 \dots\}$$

$$B : \quad \{1:2 2 2 1 1 1 \dots\} + \{0:1 1 1 1 \dots\}$$

$$C : \quad \{2:2 2 1 1 1 1 \dots\} + \{0:1 1 1 1 \dots\}$$

$$D : \quad \{2:2 1 1 1 1 \dots\} + \{0:2 1 1 1 \dots\}$$

$$E : \quad \{2:1 1 1 1 1 1 \dots\} + \{0:2 2 1 1 1 \dots\}$$

$$F : \quad \{1:1 1 1 1 1 1 \dots\} + \{0:2 2 2 1 1 \dots\}$$

$$G : \quad \{1:1 1 1 1 1 1 \dots\} + \{0:1 2 2 2 1 \dots\}$$

$$C = E > D > A = G > B = F.$$

The maximum is at convergents C and E. Indeed wherever a sequence of partial quotients $\dots 1_m 2_n \dots$ is split to form θ_{k+1} and χ_k , the maximum sum will occur with the split position at either

$$1. \theta_{k+1} = \{2:2_{n-1} 1_m \dots\}, \quad \chi_k = \{0:1_m 2_n \dots\}, \text{ or}$$

$$2. \theta_{k+1} = \{2:1_m 2_n \dots\}, \quad \chi_k = \{0:2_{n-1} 1_m \dots\}.$$

Now compare the maxima in these three groupings. The differences lie in the continued fractions commencing

$$\text{Grouping 1 } 0:2 1 1, \quad \text{Grouping 2 } 0:1 2 1, \quad \text{Grouping 3 } 0:2 2 1.$$

The least of these maxima is that of Grouping 1. Therefore this defines the critical value \mathcal{A} , corresponding to the largest minimum μ of $\epsilon_k q_k^2$. This agrees with the numerical results in Tables 24 and 25 and goes some way towards explaining why in the maximal groupings the 2s appear in pairs. Indeed, there is a symmetry between the 1 and 2 values in the partial quotients, so the 1s appear in pairs too. More generally, the critical values of \mathcal{A} have repeated sequences of even numbers of 1s, and even numbers of 2s.

We can press this further. Suppose that a grouping of 1s and 2s is associated with a maximum sum $\theta_{k+1} + \chi_k$. This must be repeated throughout the tail of the continued fraction in order for that θ to be a critical value in the Lagrange-Markoff Spectrum. Indeed, any deviation from repeating this grouping will degrade θ from being critical. Think how a sieve with an irregular, uneven mesh would let through some lumps while retaining some fine particles which should pass through. By analogy the grouping must be a strictly recurring sequence throughout the infinite tail of the continued fraction. This, of course, is characteristic of θ being a quadratic surd. This is the essential reason behind the \sqrt{D} in Tables 25 and 26.

Finally, recall the points made in §2.4, §3.2.3 that if $\theta = \{a : \underline{b, c, d, \dots y, z, a}\}$ satisfies $A\theta^2 + B\theta + C = 0$ with roots $\theta_+ > 0, \theta_- < 0$, then $\chi = \{0 : \underline{z, y, x, \dots b, a}\}$ satisfies $A\chi^2 - B\chi + C = 0$ with roots χ_+, χ_- , where the algebraic conjugate roots satisfy $\theta_+ = -\chi_- \quad \theta_- = -\chi_+$. This is the form of Eq 2.3 relevant to the Lagrange-Markoff Spectrum. Both θ and χ here evaluate to quadratic surds. So if $\theta_+ = \zeta + \sqrt{\beta} > 1$, then $\chi_+ = -\zeta + \sqrt{\beta} < 1$, and $\theta_+ + \chi_+ = 2\sqrt{\beta}$. Therefore it is $\sqrt{\beta}$ which determines μ , the minimum of relative error $\epsilon_k q_k^2$. This will be explored further below.

ζ is a fraction and later we will want to refer to its numerator, since it turns out that the denominator is $2M$. Hence write $\zeta = \xi/(2M)$.

12.3 Markoff numbers and Markoff's equation

Now return to Table 28 and look for relations between the various numbers M . Some are Fibonacci numbers (1, 2, 5, 13, 34, 89, ...) but not all. Perhaps there is a recursion relation? I cannot know how the correct relation was first discovered, perhaps 200 years ago. Perhaps someone noted that

$$3 \times 5 - 2 = 13, \quad 3 \times 13 - 5 = 34, \quad 3 \times 34 - 13 = 89 \quad \text{and even} \quad 3 \times 2 \times 5 - 1 = 29.$$

From this realisation, it is a short step to identifying that a triple of M numbers, (u, v, w) can give rise to three other triples by the operations:

$$U(u, v, w) = (3vw - u, v, w), \quad V(u, v, w) = (u, 3wu - v, w), \quad W(u, v, w) = (u, v, 3uv - w), \quad (12.2)$$

and in this process M numbers higher up the list are generated. The process and the numbers M are named after the Russian Andrei Markoff. Here are some examples starting from (1, 2, 5).

- $U(1, 2, 5) = (29, 2, 5), \quad V(1, 2, 5) = (1, 13, 5), \quad W(1, 2, 5) = (1, 2, 1)$ showing that (1, 2, 1) is also a Markoff triple.
- $V(1, 2, 1) = (1, 1, 1)$. This has the smallest values and is the seed triple for the whole Markoff process.
- $U(29, 2, 5) = (1, 2, 5), \quad V(29, 2, 5) = (29, 433, 5), \quad W(29, 2, 5) = (29, 2, 169)$. Notice how one of the operations, U , reverts to an M triple from earlier in the process. In general $U^2(u, v, w) = (u, v, w)$, and similarly for V, W .
- $U(1, 13, 5) = (194, 13, 5), \quad V(1, 13, 5) = (1, 2, 5), \quad W(1, 13, 5) = (1, 13, 34)$.

Markoff triples are characterised by their largest number; the triple in which a given M is largest is unique. The first 20 Markoff triples are listed in Table 29. Note that all odd Markoff numbers are congruent to 1 mod 4, and all even ones congruent to 2 mod 32. Also, in any triple the numbers are pairwise coprime. Markoff numbers are also often presented in a tree diagram as shown in Figure 21. The tree extends indefinitely.

1	1	1	34	13	1	433	29	5	2897	194	5
2	1	1	89	34	1	610	233	1	4181	1597	1
5	2	1	169	29	2	985	169	2	5741	985	2
13	5	1	194	13	5	1325	34	13	6466	433	5
29	5	2	233	89	1	1597	610	1	7561	194	13

Table 29: The first 20 Markoff triples.

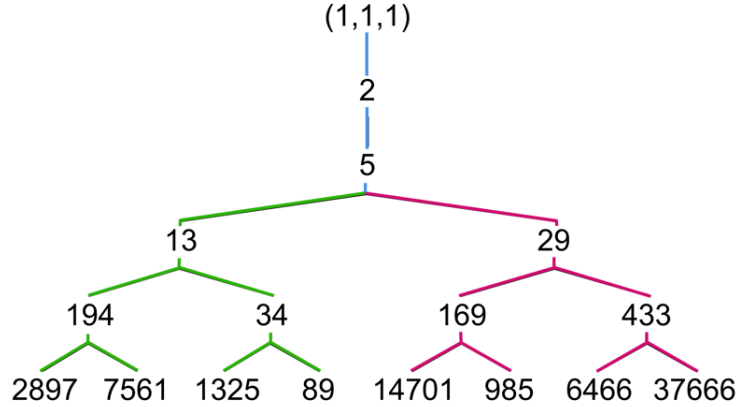


Figure 21: The Markoff numbers arranged as a tree in which new numbers are generated by a triple of numbers above.

Markoff, working in St. Petersburg, may have studied this sequence after becoming interested in Diophantine equations. He certainly knew that these triples have the following property. Sum the squares of any triple and of a triple generated from it through one of the operations U , V , W .

$$u^2 + v^2 + w^2, \text{ and } u^2 + v^2 + (3uv - w)^2.$$

Now the latter is $u^2 + v^2 + w^2 + 3uv(3uv - w) - 3uvw$.

If $u^2 + v^2 + w^2 = 3uvw$ as it is for the seed triple $(1, 1, 1)$, then $u^2 + v^2 + (3uv - w)^2 = 3uv(3uv - w)$, and the right side of each equation is in each case 3 times the product of the respective triple. In other words, each Markoff triple is a solution set of Markoff's Diophantine equation

$$u^2 + v^2 + w^2 = 3uvw. \quad (12.3)$$

Much has been written about these numbers since Markoff's two original papers of 1879 and 1880. The equation is unusual in that it has an infinity of solutions whereas other equations of the form $x^2 + y^2 + z^2 = Nxyz$, $N \neq 1$ or 3 , have no non-trivial solutions.

12.4 Indefinite binary quadratic forms

Building on early results by Hermite and by Korkine and Zolotareff, Markov saw a parallel between finding the maximum of the sum $\theta_{k+1} + \chi_k$, Eq 2.3, which defines \mathcal{A} , and finding the minimum of a corresponding binary quadratic form over the integers. The next two sub-sections look into this.

Binary quadratic forms $f(x, y) = ax^2 + 2bxy + cy^2$ were introduced in §11.2. These are usually abbreviate to (a, b, c) but I will also write them as $(a, (2b)/2, c)$ to emphasise the 2 multiplying xy .

In §11.2 we had positive definite forms, taking only positive values. Here the relevant forms are ‘indefinite’, meaning that they can evaluate both to positive and negative numbers. In his papers of 1879 and 1880 Markoff derived his spectrum by considering these indefinite forms. It will be useful to become better acquainted with these before seeing, in §12.5, how Markoff applied them to rational approximation.

An indefinite binary quadratic form, being able to take both positive and negative values, has positive discriminant $D = b^2 - ac$. D is always taken not to be a perfect square, for otherwise $f(x, y) = 0$ for some integer ratios x/y , and these cases are not interesting. Much analysis has gone into finding which integer pairs of (x, y) , if any, will make a given form (a, b, c) equal to a specified integer. A closely related question has been how to determine the minimum absolute value μ of (a, b, c) as (x, y) range over all integer pairs, excluding the trivial pair $(0, 0)$. We might expect the smallest values of (a, b, c) to cluster where x and y are small. Of course $(a, b, c) = a$ at $(\pm 1, 0)$, and $= c$ at $(0, \pm 1)$. But smaller absolute values may in some cases be attained, and many forms do indeed achieve non-trivial minima¹⁵ of either 1 or -1 . Here are some examples:

$$|(3, 4, -5)| = |(3, 8/2, -5)| = 1 \text{ at } (1, 2) \text{ and of course at } (-1, -2).$$

$$|(5, 9/2, 3)| = |5x^2 + 9xy + 3y^2| = 1 \text{ at } (\pm 1, \pm 1) \text{ and also at } (1, -2), (4, -3), (4, -9), (19, -14).$$

$$|(67, 97, 140)| = 1 \text{ at } (3, -2).$$

However others have a larger minimum, and/or at values of (x, y) removed from $(0, 0)$:

$$|(23, 78/2, 11)| \text{ has minimum } 7 \text{ at } (13, -4),$$

$$|(29, 56/2, -13)| \text{ has minimum } 8 \text{ at } (15, -7).$$

Markoff normalised the minimum values μ by dividing by \sqrt{D} . If x and y are scaled by a constant K , this ratio remains unchanged. This is because $f(Kx, Ky) = K^2 f(x, y)$ and $D(Kx, Ky) = K^4(b^2 - ac)$. Critical values of μ/\sqrt{D} are the Lagrange-Markoff ratios in Eq 12.1 and Table 28.

As with the positive definite forms of §11.2, indefinite forms can be transformed by action of a unimodular matrix, Eq 11.3, into equivalent ones, and so reduced. In geometrical terms this means changing the two basis vectors of the lattice of integer points (x, y) at which $f(x, y)$ is evaluated. Such a transformation does not alter the points which can be reached, only the path taken across the $x - y$ plane. Therefore equivalent forms have the same discriminant, determinant, and the same minimum value over the integers. The basic property of a reduced form is that, roughly speaking, the sum of the absolute values of the coefficients be small.

In §183 of his *Disquisitiones* Gauss explains one method, similar to that in §11.2, for reducing $f(x, y) = (a, b, c)$ to $\mathcal{F}(x', y') = (A, B, C)$ in which

$$0 < B < \sqrt{D}, \quad \text{and} \quad \sqrt{D} - B < |A| < \sqrt{D} + B.$$

Let $B \equiv -b \pmod{c}$ such that B lies in the interval from $\sqrt{D} - |c|$ to \sqrt{D} . Next, let $a' = (B^2 - D)/c$. If $a' < c$, the (partially) reduced form is $(c, B, a') = (A, B, C)$. If this is not fully reduced according to the above criteria, repeat the process. Eventually $a' > c$ at which point the form is fully reduced. Gauss

¹⁵‘Minimum’ here means the least absolute value for $x, y \in \mathbb{Z}$, which is of course not usually the absolute minimum over $x, y \in \mathbb{R}$.

gives the example of reducing (67, 97, 140) through (140, -97, 67) → (67, -37, 20) → (20, -3, -1) → (-1, 5, 4). His method is equivalent to building in stages the transformation matrix

$$\left(\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \right)^3 \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -3 & 4 \\ 2 & -3 \end{pmatrix}.$$

Related algorithms are described in the literature. Here is my own account which draws out some relations between binary quadratic forms, Pell's equation of §6, and continued fractions. Following Eqs 11.2, 11.3 we are looking for a unimodular transformation which will give a form equivalent to (a, b, c) but with small coefficients. So we want to find suitable u, v, u', v' in the matrix expression

$$\mathcal{F} = (X \ Y) \begin{pmatrix} u & v \\ u' & v' \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} u & u' \\ v & v' \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}, \quad uv' - u'v = 1. \quad (12.4)$$

$$\mathcal{F} = (au^2 + 2buv + cv^2)X^2 + 2(auu' + b(uv' + u'v) + cvv')XY + (au'^2 + 2bu'v' + cv'^2)Y^2.$$

The 'trick' now is to complete the square on the coefficients of X^2 and Y^2 :

$$\frac{1}{a}(a^2u^2 + 2abuv + acv^2) = \frac{1}{a}[(au + bv)^2 - v^2(b^2 - ac)] = \frac{1}{a}[(au + bv)^2 - Dv^2], \quad (12.5)$$

and similarly for Y^2 , where $D = b^2 - ac$ is the discriminant. Now we are in luck! This looks much like Pell's equation, $P^2 - DQ^2 = \kappa$, Eq 6.1. We know how to make these coefficients small, down to ± 1 ; just use the convergents of \sqrt{D} !

Example This is the example by Gauss quoted above. $f(x, y) = (67, 97, 140)$ has $D = 29$ and minimum 1 (see examples listed above). $\sqrt{29} = \{5 : 2, 1, 1, 2, 10\}$ has convergents

$$\frac{5}{1}, \quad \frac{11}{2}, \quad \frac{16}{3}, \quad \frac{27}{5}, \quad \frac{70}{13}, \quad \frac{727}{135}, \quad \frac{1524}{283}, \quad \text{etc.}$$

The sequence for κ in Pell's equation is -4, 5, -5, 4, -1, 4, -5, etc. Identify $\pm(au + bv)$ with a numerator and $\pm v$ with a denominator; these determine u . Thus $C_3 = 27/5$ leads to either $v = 5$, $u = (27 - 5.97)/67 = -458/67$, or to $v = -5$, $u = 512/67$.

Now we would like both the coefficients of X^2 and Y^2 to be as small as possible, but this is constrained by $uv' - vu' = 1$. To find which combinations are possible, construct a table of products uv' , $u'v$ corresponding to the first few convergents – see Figure 21. The denominator of u, u' , being 67, is understood. The coloured cells pick out pairs which differ by only 1, meaning that for that pair $uv' - vu' = 1$. There are 6 such coloured pairs in each box, and each will give a partially reduced form of the original form (67, 97, 140).

The reduced forms for these pairs are listed in Table 29. I will illustrate one calculation (the last pair) in full. Pair $-6655 - (-6656) = 1$. $u = 1331/67$, $v' = -5$, $u' = 512/67$, $v = -13$. Then $au + bv = 70$, and $70^2 - 29.(-13)^2 = -1$; also $au' + bv' = 27$, and $27^2 - 29.(-5)^2 = 4$. Substituting into Eq 11.12 gives

$$\mathcal{F} = \frac{1}{67}(-X^2 + 10XY + 4Y^2).$$

The factor $1/67$ equates to the $1/a$ in Eq 11.13. It is a weakness of my explanation that it produces a multiple of the given form, the factor of $1/67$ having to be disregarded. Apart from this, however, it does give the fully reduced form $(-1, 5, 4)$, the same as Gauss – a comforting thought.

u, u'	-92	-183	-275	-458	-1191	102	205	307	512	1331
v, v'										
-1						-102	-205	-307	-512	-1331
-2						-204	-410	-614	-1024	-2662
-3						-306	-615	-921	-1536	-3993
-5						-510	-1025	-1535	-2560	-6655
-13						-1326	-2665	-3991	-6656	-17303
1	-92	-183	-275	-458	-1191					
2	-184	-366	-550	-916	-2382					
3	-276	-549	-825	-1374	-3573					
5	-460	-915	-1375	-2290	-5955					
13	-1196	-2379	-3575	-5954	-15483					

Figure 22: Values of uv' and $u'v$ derived from the convergents of $\sqrt{29}$ for the case $a = 67$, $2b = 194$, $c = 140$. Coloured pairs give $uv' - vu' = 1$. The factor of $1/67$ in u, u' is omitted.

pair	$67u$	v'	$67u'$	v	form
(-183, -184)	-183	1	-92	2	(5, -3, -4)
(-275, -276)	-275	1	-92	3	(-5, -7, -4)
(-549, -550)	-183	3	-275	2	(5, 2, -5)
(-915, -916)	-183	5	-458	2	(5, 7, 4)
(-1374, -1375)	-458	3	-275	5	(4, -3, -5)
(-5954, -5955)	-458	13	-1191	5	(4, 5, -1)
(-204, -205)	102	-2	205	-1	(-4, -3, 5)
(-306, -307)	102	-3	307	-1	(-4, -7, -5)
(-614, -615)	307	-2	205	-3	(-5, 2, 5)
(-1024, -1025)	512	-2	205	-5	(4, 7, 5)
(-1535, -1536)	307	-5	512	-3	(-5, -3, 4)
(-6655, -6656)	1331	-5	512	-13	(-1, 5, 4)

Table 30: Reductions of the quadratic form (67, 97, 140) corresponding to coloured pairs in Figure 21.

In Table 29 note that the forms in the bottom panel, with v, v' negative, are associated with those in the upper panel through the coefficients a and c being interchanged. Note further that the values of a and c are equal to the values of κ from the corresponding Pell equation. This is why the most reduced value, -1 , does not appear in the form until convergent $C_4 = 70/13$ is involved: $70^2 - 29 \cdot 13^2 = -1$. Moreover, we know from §6 on Pell's equation that this pattern of κ values will repeat, so the forms themselves must repeat through a cycle matched to the recurring sequence of partial quotients. Convergent $C_9 = 9801/1820$ is the first to give $\kappa = +1$. Not every form in Table 29 is fully reduced; those with a 7 can be reduced further. For example $\begin{pmatrix} 0 & 1 \\ -1 & -3 \end{pmatrix}$ will reduce (5, 7, 4) to (4, 5, -1). In a reduced indefinite form the coefficients a and c have opposite signs.

Indefinite quadratic forms differ sharply from positive definite forms in that there is not a unique reduced form, but instead a finite number of reduced forms which are equivalent to each other. For example, if (a, b, c) is reduced, so is $(-a, b, -c)$. If the reduction algorithm is replied

repeatedly to a form which is already reduced, companion reduced forms are produced in a cycle. In the above example of (65, 97, 140), when fully reduced there are 10 independent forms amongst the 12 of Table 29. These can be placed end-to-end like dominos,

$$\begin{aligned} &(-1, 5, 4) \rightarrow (4, 3, -5) \rightarrow (-5, 2, 5) \rightarrow (5, 3, -4) \rightarrow (-4, 5, 1) \rightarrow \\ &\rightarrow (1, 5, -4) \rightarrow (-4, 3, 5) \rightarrow (5, 2, -5) \rightarrow (-5, 3, 4) \rightarrow (4, 5, -1) \text{ and back to } (-1, 5, 4), \end{aligned}$$

to make a cycle of equivalent forms. Each can be obtained from the previous by a proper equivalence transformation. For example, $\begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$ converts (4, 3, -5) to (-5, 2, 5). We previously saw palindromic symmetry in §3.4. The cyclic nature derives from the recursion sequence of partial quotients in the continued fraction for \sqrt{D} . If the recursion length L is odd, as is 5 in this example (2, 1, 1, 2, 10), the cycle length must be $2L = 10$, the 2 allowing for the alternating sign of the error in successive convergents. Conversely, for an even L , the cycle length of forms will itself be L .

It is not our main purpose here to explore indefinite binary quadratic forms further, but rather to relate them to Markoff's approach to rational approximation. He was concerned with the largest values of their minima μ normalised to \sqrt{D} . The property of indefinite forms which was crucial to his purpose is that the minimum μ is the least absolute value which appears as coefficient a (or c) within all equivalent reduced forms over the whole cycle. In his 1880 paper Markoff defined a binary quadratic form related to both the continued fractions on §12.4 and the Markoff equation of §12.3. Several of the subsequent textbooks and monographs state Markoff's quadratic forms without explaining simply how they come about. Here I attempt some rationale.

12.5 Markoff forms

Markoff constructed a family of indefinite binary quadratic forms corresponding to his numbers M and the sequences of partial quotients $1_2 2_2$, $1_2 2_4$ etc. These forms are closely related to the quadratic equations satisfied by the recurring continued fractions in Tables 27 and more fully in Table 31. For example if $x = \{0 : \underline{1_4} \ 2_2\}$, then $x = \{0 : 1_4, 2_2, x\}$. Evaluating the continued fraction,

$$x = \frac{8x + 19}{13x + 31} \quad \text{so} \quad 13x^2 + 23x - 19 = 0.$$

The corresponding Markoff form would be $13x^2 + 23xy - 19y^2$.

To put this on a more structured footing, consider a quadratic form factorised into two linear forms:

$$ax^2 + 2bxy + cy^2 = a(x - \theta_+ y)(x - \theta_- y) \tag{12.6}$$

where $\theta_+ = \zeta + \sqrt{\beta}$, $\theta_- = \zeta - \sqrt{\beta}$ are the roots of $ax^2 + 2bx + c = 0$. Because a and c have opposite signs in a reduced indefinite form, so do these two roots. If $\theta_+ > 1$ is identified with θ_{k+1} in Eq 2.3, $\theta_- < 1$ can be identified with $-\chi_k$ because the algebraic conjugate roots satisfy $\theta_+ = -\chi_-$, $\theta_- = -\chi_+$. Eq 12.6 becomes $a(x - \theta_{k+1}y)(x + \chi_k y)$.

Markoff chose θ_{k+1} , χ_k such that

- each θ_{k+1} and χ_k pair come from a critical value of θ in Table 27,
- the split in the sequences of partial quotients of 1s and 2s is made, as in §11.4.2, to give the maximum in $\theta_{k+1} + \chi_k$
- the critical sequences of 1s and 2s are those corresponding to the largest minimum μ .

Markoff took $a = M > 0$, a Markoff number. Then the Markoff form corresponding to a critical θ is

$$a \left(x - [\zeta + \sqrt{\beta}]y \right) \left(x + [-\zeta + \sqrt{\beta}]y \right) = Mx^2 - 2M\zeta xy + M(\zeta^2 - \beta)y^2. \quad (12.7)$$

Subject to these criteria $\theta_+ + \chi_+ = 2\sqrt{\beta} = \mathcal{A}$. Since $\theta_{k+1} > 1$ and $\chi_k < 1$, the coefficient $b < 0$, and in his 1879 paper all coefficients b are indeed negative. More recent accounts of his work, such as the monographs by Cusick and Flahive, and by Cassels (see Bibliography at the beginning of this article), record b as positive, corresponding to factorising Eq 12.6 as $a(x - \chi_+y)(x - \chi_-y)$. These companion forms have the same discriminant and determinant, and the choice does not seem to affect the conclusions.

The Lagrange-Markoff spectrum, the Markoff equation and Markoff forms have patterns and inter-relations which have held the interest of many mathematicians. I will further develop this account of mine in steady stages, continuing now with some examples of Markoff forms, before looking at some general expressions.

Examples

1. The first critical value corresponds to $a_k = \dots 1 \ 1 \ 1 \ 1 \ 1 \ 1 \dots$. This splits as $\theta_{k+1} = \{1 : \underline{1}\} = G = \frac{1}{2}(1 + \sqrt{5})$, $\chi_k = \{0 : \underline{1}\} = 1/G$. The Markoff form is $a(x - Gy)(x + y/G) = a(x^2 - xy - y^2)$.
2. The second critical value corresponds to the sequence $\dots 2 \ 2 \ 2 \ 2 \ 2 \ 2 \dots$. This splits as $\theta_{k+1} = \{2 : \underline{2}\} = 1 + \sqrt{2}$, $\chi_k = \{0 : \underline{2}\} = -1 + \sqrt{2}$. The Markoff form is

$$a \left(x - (1 + \sqrt{2})y \right) \left(x + (-1 + \sqrt{2})y \right) = a(x^2 - 2xy - y^2).$$

If $a = 2 = M$, a Markoff number, the form is $2x^2 - 4xy - 2y^2$. In Cusick and Flahive this appears as $2x^2 + 4xy - 2y^2$.

3. The recurring sequence $\dots 1 \ 1 \ 2 \ 2 \ 1 \ 1 \ 2 \ 2 \ 1 \ 1 \dots$ gives its largest value of $\theta_{k+1} + \chi_k$ when split into $\theta_{k+1} = \{2 : \underline{1 \ 1 \ 2 \ 2}\}$, $\chi_k = \{0 : \underline{2 \ 1 \ 1 \ 2}\}$. One continued fraction is shifted one place relative to the other, so $\chi_k = 1/\theta_{k+1}$. If again a is taken as the current Markoff number, now 5, the form is

$$5 \left(x - \frac{1}{10}(11 + \sqrt{221})y \right) \left(x + \frac{1}{10}(-11 + \sqrt{221})y \right) = 5x^2 - 11xy - 5y^2.$$

as obtained by Markoff.

4. The sequence $\dots 1 \ 1 \ 2 \ 2 \ 1 \ 1 \ 2 \ 2 \ 1 \ 1 \dots$ in item 3 above can alternatively be split into $\theta_{k+1} = \{2 : \underline{2 \ 1 \ 1 \ 2}\}$, $\chi_k = \{0 : \underline{1 \ 1 \ 2 \ 2}\}$. This gives the same sum since the integer parts have merely been exchanged. Here $\theta_{k+1} = (9 + \sqrt{221})/10$ and χ_k is again the negative of its conjugate. Now the Markoff form is $(a/5)(5x^2 - 9xy - 7y^2)$. This looks quite different, but is equivalent to $5x^2 + 11xy - 5y^2$, the form given by Cusick and Flahive. The matrix $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ converts $5x^2 - 9xy - 7y^2$ to $5x^2 + 11xy - 5y^2$. $5x^2 + 11xy - 5y^2$ is not properly equivalent to $5x^2 - 11xy - 5y^2$, but the fact that both forms can be obtained from the same sequence of partial quotients emphasises their essential compatibility.
5. The recurring sequence $\dots 1_2 \ 2_4 \dots$ was illustrated in Figure 17. As with the cases 3 and 4 above, this can be split in two ways, each of which gives the maximum value of the sum $\theta_{k+1} + \chi_k$:
Split 1: $\{2 : \underline{1_2 \ 2_4 \dots}\} + \{0 : \underline{2 \ 2 \ 2 \ 1_2 \ 2_4 \dots}\}$. Here $\theta_{k+1} = \frac{1}{58}(63 + \sqrt{7565}) = 2 \cdot 58581$ and $\chi_k = \frac{1}{58}(-63 + \sqrt{7565}) = 0 \cdot 41340$. The Markoff form is now $29x^2 - 63xy - 31y^2$, which is as

Markoff himself stated.

Split 2: $\{ 2: 2 \ 2 \ 2 \ \underline{1_2} \ \underline{2_4} \ \dots \} + \{ 0: \underline{1_2} \ \underline{2_4} \ \dots \}$. Here $\theta_{k+1} = \frac{1}{58}(53 + \sqrt{7565}) = 2 \cdot 41340$ and $\chi_k = \frac{1}{58}(-53 + \sqrt{7565}) = 0 \cdot 58581$. The Markoff form is therefore $29x^2 - 53xy - 41y^2$.

The form from Split 2 is transformed by $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ to $29x^2 + 63xy - 31y^2$ as quoted by Cusick and Flahive.

The next step in developing the Markoff form is to recall that the Lagrange-Markoff ratio in Eq 12.1 is $1/\mathcal{A} = M/\sqrt{D}$ where $D = 9M^2 - 4$. Above it was also shown that $\mathcal{A} = 2\sqrt{\beta}$. Moreover, the discriminant $\Delta = 'b^2 - 4ac'$ of Eq 12.7 is $4M^2\beta$, making $\beta = \Delta/(4M^2)$. Putting these together shows that $\Delta = D$ and $\beta = (9M^2 - 4)/(4M^2)$. It is necessary to take M to be the highest Markoff number in the triple corresponding to the form. Subject to this, the general structure of the Markoff form simplifies to

$$Mx^2 - 2M\zeta xy + \frac{1}{4M}(4M^2\zeta^2 - 9M^2 + 4)y^2. \quad (12.8a)$$

Since the fraction $\zeta = \xi/(2M)$ this can be written alternatively in terms of the numerator ξ :

$$Mx^2 - \xi xy + \frac{1}{4M}(\xi^2 - 9M^2 + 4)y^2. \quad (12.8b)$$

The link between Markoff's quadratic form and Markoff's equation, Eq 12.3, is made more clear by writing both in completed squares form. The form Eq 12.8 is

$$M[(x - \zeta y)^2 - \frac{1}{4M^2}(9M^2 - 4)y^2] = \frac{1}{4M}[(2M(x - \zeta y))^2 - (9M^2 - 4)y^2]. \quad (12.9)$$

The equivalent operation with Markoff's equation is to assume that $u \leq v \leq w$, multiply by 4 to clear fractions and relabel w as M , the largest Markoff number. This gives

$$4u^2 + 4v^2 + 4M^2 - 12Muv = 0.$$

Now complete the square:

$$(2u - 3Mv)^2 = 4u^2 - 12Muv + 9M^2v^2 \quad \text{so} \quad (2u - 3Mv)^2 - 9M^2v^2 + 4v^2 = -4M^2.$$

Finally write $2u - 3Mv = X$

$$\frac{1}{4M}[X^2 - Nv^2] = -M, \quad N = 9M^2 - 4 \quad (12.10)$$

in which X , v and N are integers. Notice how this looks not only like Eq 12.9 if $y = v$, but also like Pell's equation. However, I don't think the analogy can be pushed any further.

Bear in mind that though Eq 12.8 is a general expression for the Markoff form for M , it is not unique; there is a cycle of equivalents. However, a Markoff form always has a Markoff number as coefficient a so that the minimum value, μ , of the form is M , obtained for $x = 1$, $y = 0$. We saw in Examples 3, 4, 5 above that there are (at least) two ways of splitting the sequence of partial quotients ... $1 \ 1 \ 2 \ 2 \ 1 \ 1 \ \dots$ etc, to produce a maximum value of $\theta_{k+1} + \chi_k$, these corresponding to swapping the position of the integer 2. Suppose s and t represent forward and reversed sequences of partial quotients excluding the integer part. Then

$$\textit{Split 1} : \quad \theta_{k+1} = \{2 : s\}, \quad \chi_k = \{0 : t\},$$

$$\textit{Split 2} : \quad \theta_{k+1} = \{2 : t\}, \quad \chi_k = \{0 : s\},$$

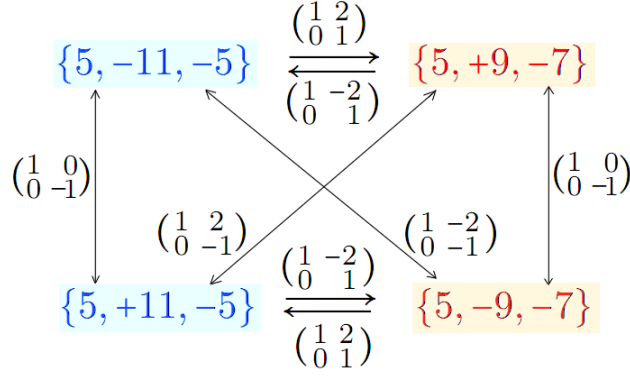


Figure 23: Transformations of Split 1 and Split 2 versions of a Markoff form under actions of proper and improper equivalence matrices.
Illustration for case $M = 5$

Now if Split 1 evaluates as $\theta_{k+1} = \zeta + \sqrt{\beta}$, $\chi_k = -\zeta + \sqrt{\beta}$, these being negative algebraic conjugates, then Split 2 must evaluate as $\theta_{k+1} = \zeta - 2 + \sqrt{\beta}$, $\chi_k = 2 - \zeta + \sqrt{\beta}$. The quadratic forms are obtained from Eq 12.6:

$$\text{Split 1: } M[x^2 - 2\zeta xy + (\zeta^2 - \beta)y^2], \quad \zeta = \frac{\xi}{2M}, \beta = \frac{9M^2 - 4}{4M^2},$$

$$\text{Split 2: } M[x^2 + 2(\zeta - 2)xy + (\zeta^2 - \beta - 4\zeta + 4)y^2],$$

In matrix form these two versions of a Markoff form are equivalent through a proper transformation:

$$\begin{pmatrix} 1 & \zeta \\ \zeta & \zeta^2 - \beta \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} x & \zeta - 2 \\ \zeta - 2 & (\zeta^2 - \beta - 4\zeta + 4) \end{pmatrix}. \quad (12.11)$$

Table 31 lists both versions of forms for all Markoff numbers $< 2,000,000$. n merely indexes the forms in increasing order. ξ is the numerator of the fractional parts of θ_{k+1} , χ_k . The form according to Split 1 is not listed explicitly because $a = -c = M$, $b = -\xi$. For each M the Split 1 and Split 2 versions of the Markoff form are related through proper or improper equivalence transformation as shown in Figure 23, which illustrates the case $M = 5$. (Note that here the triple denotes the actual coefficients.) Since all indefinite reduced forms will have several equivalent ones, there is no definitive Markoff form, though every Markoff form has $a = M$. M is the lowest value for a of all equivalent forms in the cycle, this occurring at $(x, y) = (1, 0)$. That is why these forms were Markoff's tool for investigating the critical minima of $\epsilon_k q_k^2$.

There is a crucial relation between a Markoff triple $\{M_1, M_2, M_3\}$ and the recurring sequences of partial quotients representing M_1, M_2, M_3 . If M_3 is the largest and represented by sequence S_3 , and similarly for M_1 and M_2 , then

$$S_3 = S_1 S_2 = S_2 S_1 \text{ by concatenation of strings.} \quad (12.12)$$

This remarkable fact can be observed throughout Table 31. Here are two examples for triples listed in Table 29:

1. $\{2, 169, 985\} : 2 \equiv 2_2$ and $169 \equiv 1_2 2_6$ so $985 \equiv 1_2 2_6 2_2 = 1_2 2_8$.
2. $\{13, 194, 7561\} : 13 \equiv 1_4 2_2$ and $194 \equiv 1_2 (1_2 2_2)_2$ so $985 \equiv 1_2 (1_2 2_2)_2 1_4 2_2 = 1_4 2_2 1_2 2_2 1_4 2_2 = (1_4 2_2)_2 1_2 2_2$, since the sequence repeats indefinitely.

n	M	ξ	cont frac	Split 2		
			sequence	$a = M$	b	c
1	1	1	1_2			
2	2	4	2_2			
3	5	11	$1_2 2_2$	5	9	-7
4	13	29	$1_4 2_2$	13	23	-19
5	29	53	$1_2 2_4$	29	63	-31
6	34	76	$1_6 2_2$	34	60	-50
7	89	199	$1_8 2_2$	89	157	-131
8	169	309	$1_2 2_6$	169	367	-181
9	194	344	$1_2 (1_2 2_2)_2$	194	432	-196
10	233	521	$1_{10} 2_2$	233	411	-343
11	433	791	$(1_2 2_2)_2 2_2$	433	941	-463
12	610	1364	$1_{12} 2_2$	610	1076	-898
13	985	1801	$1_2 2_8$	985	2139	1055
14	1325	2339	$1_2 (1_4 2_2)_2$	1325	2961	-1327
15	1597	3571	$1_{14} 2_2$	1597	2817	-2351
16	2897	5137	$1_2 (1_2 2_2)_3$	2897	6451	-2927
17	4181	9349	$1_{16} 2_2$	4181	7375	-6155
18	5741	10497	$1_2 2_{10}$	5741	12467	-6149
19	6466	11812	$(1_2 2_2)_3 2_2$	6466	14052	-6914
20	7561	13407	$(1_4 2_2)_2 1_2 2_2$	7561	16837	-7639
21	9077	16013	$1_2 (1_6 2_2)_2$	9077	20295	-9079
22	10946	24476	$1_{18} 2_2$	10946	19308	-16114
23	14701	26879	$(1_2 2_4)_2 2_2$	14701	31925	-15745
24	28657	64079	$1_{20} 2_2$	28657	50549	-42187
25	33461	61181	$1_2 2_{12}$	33461	72663	-35839
26	37666	68808	$1_2 2_2 (1_2 2_4)_2$	37666	81856	-40276
27	43261	76711	$1_2 (1_2 2_2)_4$	43261	96333	-43709
28	51641	91161	$1_2 (1_4 2_2)_3$	51641	115403	-51719
29	62210	109736	$1_2 (1_8 2_2)_2$	62210	139104	-62212
30	75025	167761	$1_{22} 2_2$	75025	132339	-110447
31	96557	176389	$(1_2 2_2)_4 2_2$	96557	209839	-103247
32	135137	238555	$(1_6 2_2)_2 1_4 2_2$	135137	301993	-135341
33	195025	356589	$1_2 2_{14}$	195025	423511	-208885
34	196418	439204	$1_{24} 2_2$	196418	346468	-289154
35	294685	522529	$(1_4 2_2)_3 1_2 2_2$	294685	656211	-297725
36	426389	752123	$1_2 (1_{10} 2_2)_2$	426389	953433	-426391
37	499393	913103	$(1_2 2_6)_2 2_2$	499393	1084469	-534883
38	514229	1149851	$1_{26} 2_2$	514229	907065	-757015
39	646018	1145528	$1_2 (1_2 2_2)_5$	646018	1438544	-652708
40	925765	1633169	$1_2 (1_6 2_2)_3$	925765	2069891	-925969
41	1136689	2078353	$1_2 2_{16}$	1136689	2468403	-1217471
42	1278818	2338164	$(1_2 2_4)_3 2_2$	1278818	2777108	-1369634
43	1441889	2634023	$(1_2 2_2)_5 2_2$	1441889	3133533	-1541791
44	1686049	2989713	$1_2 (1_2 2_2)_3 \dots$ $\dots 1_2 (1_2 2_2)_2$	1686059	3744523	-1703441

Table 31: Markoff forms for $M < 2,000,000$.

I can offer the following partial explanation for this. In general, for any continued fraction representing $\theta < 0$, if ρ is the tail, then $\theta = \{0 : a_1, a_2, \dots, a_k + \rho\} = \frac{A\rho + C}{B\rho + D}$ where C/D is the convergent C_k and A/B the previous convergent C_{k-1} . If a_1, a_2, \dots, a_k is a recurring sequence, then $\rho = \{0 : a_1, a_2, \dots, a_k + \rho\}$. Suppose there is also a second recurring sequence satisfying $\sigma = \{0 : b_1, b_2, \dots, b_h + \sigma\}$ where $\sigma = \frac{P\sigma + R}{Q\sigma + S}$. This rational equation immediately evaluates to the quadratic $Q\sigma^2 - (P - S)\sigma - R = 0$; note how the quadratic is recognised within the coefficients of the rational function. Concatenating these two sequences into the recurring $\tau = a_1, a_2, \dots, a_k, b_1, b_2, \dots, b_h$ is equivalent to substituting $\rho = \frac{P\sigma + R}{Q\sigma + S}$ into the right-hand side of the first fraction. This gives

$$\tau = \frac{\tau(AP + BR) + (CP + DR)}{\tau(AQ + BS) + (CQ + DS)},$$

$$\text{with } \tau^2(AQ + BS) - \sigma(AP + BR - CQ - DS) - (CP + DR) = 0.$$

The alternative substitution of $\sigma = \frac{A\rho + C}{B\rho + D}$ into the second fraction gives

$$v = \frac{v(AP + CQ) + (AR + CS)}{v(BP + DQ) + (BR + DS)},$$

$$\text{with } v^2(BP + DQ) - v(AP + BR + CQ - DS) - (AR + CS) = 0$$

being the quadratic equation. Now apply this generality to the specific case of Markoff continued fractions. The relations between the convergents A/B , C/D , P/Q , R/S are special, coming from the critical continued fractions in the Lagrange-Markoff spectrum. They are such that there is an inductive chain, beginning with the seed sequences 1_2 , 2_2 , which simultaneously produces further Markoff numbers and companion Markoff forms. The chain starts as shown below. Here s denotes the sequence 1_2 , t denotes 2_2 , $u_1 = st$ concatenated as $1_2 2_2$, $u_2 = ts = 2_2 1_2$, and similarly for w_1 , etc. For each Markoff number M the rational expression is obtained by evaluating the recurring sequence, and the coefficients of the quadratic which follows from it are shown in $\{\dots\}$.

$$M = 1 : \frac{s+1}{s+2} \rightarrow \{1, 1, -1\}, \quad M = 2 : \frac{t+2}{2t+5} \rightarrow \{2, 4, -2\},$$

$$M = 5 : \frac{3u_1 + 7}{5u_1 + 12} \rightarrow \{5, 9, -7\}, \quad \frac{3u_2 + 5}{7u_2 + 12} \rightarrow \{7, 9, -5\}$$

$\{5, 9, -7\}$ and $\{7, 9, -5\}$ are both equivalent to the standard Markoff form $\{5, 11, -5\}$. The two expressions for $M = 5$ represent the Markoff triple $(1, 2, 5)$ because the '5' is generated by concatenation of the '1' and '2' cases. The chain continues by concatenating the '1' and '5' cases:

$$M = 13 : \frac{8w_3 + 19}{13w_3 + 31} \rightarrow \{13, 23, -19\}, \quad \frac{10w_2 + 17}{17w_2 + 29} \rightarrow \{17, 19, -17\},$$

$$\frac{8w_1 + 13}{19w_1 + 31} \rightarrow \{19, 23, -13\}.$$

The three versions correspond to different substitutions which themselves correspond to concatenations $1_2 1_2 2_2$, $1_2 2_2 1_2$ and $2_2 1_2 1_2$, and all correspond to the Markoff triple $(1, 5, 13)$. Observe that the Markoff number 13 does not appear in all three rational functions in w_1 , w_2 , w_3 ; the Markoff number is the smallest coefficient of w_j over all denominators (equal to the smallest constant term over the numerators). Continuing the chain, the triple $(2, 5, 29)$ follows from substituting $\frac{t+2}{2t+5}$ into the two expressions for $M = 5$ and so generates the equivalent forms with coefficients $\{29, 53, -41\}$, $\{31, 61, -31\}$, $\{-29, 63, 31\}$. Clearly as M increases so does the number of possible ways of substituting, and thus the number of equivalent quadratic forms. However M can always be determined from

any version of the quadratic form because all versions have the same discriminant, which equates to $9M^2 - 4$. This is a cumbersome expression. However $AD - BC = PS - QR = +1$, because the last convergent is always of even index ($C_2, C_4, C_6, \text{etc}$) since the recurring sequences all have an even number of partial quotients. The expression for $9M^2$ is then a perfect square and $3M$ boils down to the pleasingly simple expression

$$3M = AP + BR + CQ + DS. \tag{12.13}$$

This is the scalar (dot) product of the two vectors (A, B, C, D) and (P, Q, R, S) .

The Splits 1 and 2 which produce the maximum of $\theta_{k+1} + \chi_k$ are only two of the possible positions at which the sequence of recurring partial quotients can be cut. A split is possible between every adjacent pair of partial quotients, and evaluation of these according to Eq 12.6 generates a cycle of forms whose length L equals the number of partial quotients in the recurring sequence, as we saw in §12.4. There are, however, two complications to this:

- the forms produced direct from Eq 12.6 will not necessarily be reduced according to the Gauss criteria of §12.4,
- Gauss's algorithm for producing a cycle of forms from a given reduced form will not in general generate successive forms in the same order as the continued fraction.

Table 32 illustrates this for the case of $M = 433$ with sequence $1_2 2_2 1_2 2_4$. The ten split positions are labelled A to J as follows:

$$\dots 1_A 2_B 2_C 1_D 1_E 2_F 2_G 2_H 2_I 1_J 1_A 2_B \dots$$

Thus, splitting at A gives

$$\theta = \{2 : 2 \ 1_2 \ 2_4 \ 1_2 \ 2_2 \ \dots\} = \frac{1}{874}(787 + \sqrt{D}), \quad D = 1687397,$$

$$\chi = \{0 : 1_2 \ 2_4 \ 1_2 \ 2_2 \ \dots\} = \frac{1}{874}(-787 + \sqrt{D}).$$

The numerator, ξ and denominator of the fractional part ζ (here 787 and 874 respectively) are listed in columns 2 and 3 of Table 32. Next follows the form obtained by putting these values into Eq 12.6. These are not Gauss-reduced, so the next panel of three columns gives the reduced equivalents. The right-most panel, $\{a', b', c'\}$, gives the cycle of Gauss-reduced forms obtained by starting from $\{-611, 435, 613\}$ at split position A. Though the three panels of forms are different in order and signs of the coefficients, they do all present the same information. They show how symmetry in the sequence of partial quotients gives rise to symmetry in the a and c coefficients, as in $\{463, -911, -463\}$ at Split G panel one, Split H panel two, Split D panel three.

12.5.1 Symmetric Markoff forms, $a = -c$

Two special types of form deserve comment, the first having $a = -c$. These occur under Split 1 when the recursion sequence in the continued fraction has only one pair of 2s, that is $\dots 1_m 2_2 1_m \dots$, m even. Table 33 spells out their properties. The symmetric form in question appears in the central panel with coefficients $a = M, b, c$. This is how forms appear in Markoff's 1879 paper. The last panel of three coefficients labeled A, B, C gives an equivalent under Split 2.

Notice that the Markoff numbers (column 4) are the alternate Fibonacci numbers **2, 3, 5, 8, 13, 21, 34, 55, 89**, etc. These arise from the consecutive 1s in the continued fraction; recall §3.2.1

Split point			from Eq 12.6			Gauss-reduced			Gauss cycle from A, reduced		
	num	den	a	b	c	A	B	C	a'	b'	c'
A	787	874	437	-787	-611	-611	435	613	-611	435	613
B	961	874	437	-961	-437	-437	787	611	613	791	-433
C	787	1222	611	-787	-437	-437	961	437	-433	941	463
D	435	1226	613	-435	-611	-611	787	437	463	911	-463
E	791	866	433	-791	-613	-613	435	611	-463	941	433
F	941	926	463	-941	-433	-433	791	613	433	791	-613
G	911	926	463	-911	-463	-463	941	433	-613	435	611
H	941	866	433	-941	-463	-463	911	463	611	787	-437
I	791	1226	613	-791	-433	-433	941	463	-437	961	437
J	435	1222	611	-435	-613	-613	791	433	437	787	-611

Table 32: Cycles of forms obtained by splitting the recurring sequence of partial quotients $1_2 2_2 1_2 2_4$ at various points. Three sets of cycles are presented.

n	cont fracn. sequence	ξ	Split 1			Split 2		
			Symmetric, $b = \pm\sqrt{5M^2 - 4}$			Non-symmetric Markoff		
			$a = M$	b	$c = -M$	$A = M$	$B = -\xi$	C
1	all 1	1	1	-1	-1			
2	all 2	4	2	-4	-2			
3	$1_2 2_2$	9	5	-11	-5	5	-9	-7
4	$1_4 2_2$	23	13	-29	-13	13	-23	-19
6	$1_6 2_2$	60	34	-76	-34	34	-60	-50
7	$1_8 2_2$	157	89	-199	-89	89	-157	-131
10	$1_{10} 2_2$	411	233	-521	-233	233	-411	-343
12	$1_{12} 2_2$	1076	610	-1364	-610	610	-1076	-898
15	$1_{14} 2_2$	2817	1597	-3571	-1597	1597	-2817	-2351
17	$1_{16} 2_2$	7375	4181	-9349	-4181	4181	-7375	-6155
22	$1_{18} 2_2$	19308	10946	-24476	-10946	10946	-19308	-16114
24	$1_{20} 2_2$	50549	28657	-64079	-28657	28657	-50549	-42187
30	$1_{22} 2_2$	132339	75025	-167761	-75025	75025	-132339	-110447
34	$1_{24} 2_2$	346468	196418	-439204	-196418	196418	-346468	-289154
38	$1_{26} 2_2$	907065	514229	-1149851	-514229	514229	-907065	-757015

Table 33: Symmetric Markoff forms $Mx^2 \pm bxy - My^2$. The last three columns show the equivalent non-symmetric form with $b = -\xi$ obtained under Split 2.

which showed that the continued fraction for the Golden Mean $G = \{1 : \underline{1}\}$ has convergents $\frac{3}{2}, \frac{5}{3}, \frac{8}{5}, \frac{13}{8}, \text{etc.}$ I will say more about this in the next subsection.

For these symmetrical forms it is easy to solve c in Eq 12.8 for b , giving $b^2 = 5M^2 - 4$ and the form is

$$Mx^2 - \sqrt{5M^2 - 4}xy - My^2, \quad \text{symmetric Markoff type,} \quad (12.14)$$

where the root is an integer. The values of M and b in the central panel of Table 32 are the only pairs of numbers $< 2,000,000$ for which $\sqrt{5 \times \text{integer}^2 - 4}$ is an integer, so this can be taken as a

defining characteristic. Applying $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ to $\{M, -\sqrt{5M^2-4}, -M\}$ gives the Split 2 version:

$$Mx^2 + (4M - \sqrt{5M^2-4})xy - (-3M + 2\sqrt{5M^2-4})y^2. \quad (12.15a)$$

This is a general expression for the non-symmetric forms in the right panel on Table 32 and it also gives a formula for ξ :

$$\xi = 4M - \sqrt{5M^2-4}. \quad (12.15b)$$

In the symmetric form of the central panel, as $a = M \rightarrow \infty$, so $b \rightarrow \sqrt{5}M$. In the right hand panel the corresponding asymptotic limits are $B/A \rightarrow 1 \cdot 7639 = 4 - \sqrt{5}$ and $C/A = C/M \rightarrow 1 \cdot 4721 = -3 + 2\sqrt{5}$.

12.5.2 Forms due to continued fractions of type $\underline{1_2 2_n}$

The next most simple type of Markoff form is the complement of that above, with the 1s and 2s interchanged, so they have a recurring sequence $\dots 1_2 2_n \dots$ for n even. As with the $\dots 1_m 2_2 \dots$ type, there are two equivalent Markoff forms depending on whether the continued fraction is split to give

$$\textit{Split 1: } \theta_{k+1} = \{2 : \underline{1_2 2_n}\}, \quad \chi_k = \{0 : 2_{n-1} \underline{1_2 2_n}\}, \text{ or}$$

$$\textit{Split 2: } \theta_{k+1} = \{2 : 2_{n-1} \underline{1_2 2_n}\}, \quad \chi_k = \{0 : \underline{1_2 2_n}\}.$$

Table 34 lists the recurring sequences and the forms corresponding to Split 1 and Split 2. n in column 1 indexes the Markoff numbers M . Split 1 is the form quoted by Cusick and Flahive, and by Markoff himself. Note that the coefficients a and c are more nearly equal than with Split 2. The last column is included as a mathematical curiosity; it refers to an equivalent symmetric non-Markoff form in which coefficient $b = M$, produced by splitting the sequence of partial quotients at its other symmetrical position. Bearing in mind the explanation of properly and improperly equivalent forms in Figure 23, I will not labour to distinguish the sign of coefficient b .

n	cont fracn	Split 1			Split 2			Symmetric, $b = M$.
		$a = M$	$b = \pm\xi$	c	a	b	c	$a = -c \rightarrow \sqrt{2}M$
3	$1_2 2_2$	5	11	-5	5	-9	-7	
5	$1_2 2_4$	29	63	-31	29	-53	-41	41
8	$1_2 2_6$	169	367	-181	169	-309	-239	239
13	$1_2 2_8$	985	2139	-1055	985	-1801	-1393	1393
18	$1_2 2_{10}$	5741	12467	-6149	5741	-10497	-8119	8119
25	$1_2 2_{12}$	33461	72663	-35839	33461	-61181	-47321	47321
33	$1_2 2_{14}$	195025	423511	-208885	195025	-356589	-275807	275807
41	$1_2 2_{16}$	1136689	2468403	-1217471	1136689	-2078353	-1607521	1607521

Table 34: Markoff forms of type $1_2 2_n$.

It is possible to calculate the coefficients of these Markoff forms for this type in terms of the convergents of $-1 + \sqrt{2} = \{0 : \underline{2}\}$. For this subsection only, let χ_2 denote $\{0 : \underline{1_2 2_n}\}$, the value of χ_k at Split 2 above. Similarly let $\chi_1, \theta_1, \theta_2$ be the other corresponding values. These are related through

$$\theta_1 = 2 + \chi_2, \quad \theta_2 = 2 + \chi_1,$$

$$\chi_1 = \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \dots \frac{1}{2 + \chi_2}}}}, \quad \text{with } (n-1) \text{ '2's,} \quad = \frac{P\chi_2 + Q}{Q\chi_2 + R},$$

$$\chi_2 = \frac{1}{1 + \frac{1}{1 + \frac{1}{2 + \chi_1}}} = \frac{\chi_1 + 3}{2\chi_1 + 5}.$$

Here P, Q, R are integers obtained by evaluating χ_1 as an ordinary fraction. Because of all the 2s, their values are related to the convergents of $\{0 : \underline{2}\} = -1 + \sqrt{2}$ which are

$$\frac{1}{2}, \quad \frac{2}{5}, \quad \frac{5}{12}, \quad \frac{12}{29}, \quad \frac{29}{70}, \quad \frac{70}{169}, \quad \frac{169}{408}, \quad \frac{408}{985}, \quad \dots$$

Thus

$$n = 2 \rightarrow \chi_1 = \frac{P\chi_2 + Q}{Q\chi_2 + R} = \frac{1}{\chi_2 + 2}, \quad n = 4 \rightarrow \frac{2\chi_2 + 5}{5\chi_2 + 12}, \quad n = 6 \rightarrow \frac{12\chi_2 + 29}{29\chi_2 + 70}, \quad \text{etc.}$$

The values of $(P\chi_2 + Q)/(Q\chi_2 + R)$ are readily found using the matrix notation of §1.4.1:

$$\begin{pmatrix} P\chi_2 & Q \\ Q\chi_2 & R \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}^{n-1} \begin{pmatrix} \chi_2 \\ 1 \end{pmatrix}.$$

You will notice within the above list of convergents several Markoff numbers: 1, 2, 5, 29, 169, 985, 5741, 33461, 195025, *etc.* This explains how they arise. Solving simultaneously the two expressions involving χ_1, χ_2 shows that χ_1 satisfies the quadratic

$$(Q + 2R)\chi_1^2 + (5R - P + Q)\chi_1 - (3P + 5Q) = 0, \quad (12.16a)$$

whilst χ_2 satisfies

$$(2P + 5Q)\chi_2^2 + (5R - P - Q)\chi_2 - (Q + 3R) = 0. \quad (12.16b)$$

The coefficients of these two quadratics are respectively the three coefficients of the Split 1 and Split 2 Markoff forms in Table 34 (except that in Split 2 the sign of b is reversed). Clearly $Q + 2R = 2P + 5Q = M$. The convergents of $\pm 1 + \sqrt{2}$ play the same role here as do the Fibonacci numbers in the $\underline{1}_m \underline{2}_2$ type, §12.5.1.

Looking at the limiting ratios of the two Markoff forms we see that with Split 1 the ratio $|b/a| \rightarrow 2 \cdot 1716 = 5 - \sqrt{8}$, and $|c/a| \rightarrow 1 \cdot 0711 = 5\sqrt{2} - 6$. On the other hand, with Split 2 the ratio $|b/a| \rightarrow 1 \cdot 8284 = \sqrt{8} - 1$, and the ratio $|c/a| \rightarrow 1 \cdot 4142 = \sqrt{2}$. These are readily derived from Eq 15 a, b using the limiting values $P/Q \rightarrow Q/R \rightarrow -1 + \sqrt{2}$.

12.6 Wider aspects of the Markoff spectrum

THIS SUB-SECTION IS SUBJECT TO REVISION

Mathematical interest in the Lagrange-Markoff Spectrum and the Markoff numbers has been maintained because they can be studied from several quite different points of view. Four particular have been

torus. The line L is now wrapped round the torus. If $\tan \theta$ were rational, L would close back on itself as a single loop round the torus, passing through the hole in its centre. If not, it will continue to spiral indefinitely round and through the torus. So this links to the fundamental group of the torus. The loop can be 'pulled tight' to lie on geodesics on the torus.

Harvey Cohn has shown that lines L whose cutting sequence corresponds to a Markoff continued fraction do not pass closer to the corner of the unit cell than $1/3$ of the cell width. This translates on the torus to them avoiding a patch on the torus. The $1/3$ value relates to the limit point of the Lagrange-Markoff ratios being $1/3$. Since a patch on the torus is avoided by these looped lines, that part of the torus could be removed without disturbing the loop geodesics. For this reason Markoff forms are said to be represented by geodesics on a punctured torus.

Another representation is by Fuchisan groups.

13 Liouville's theorem and transcendental numbers

NOTE: THIS SECTION IS INCOMPLETE AND SUBJECT TO REVISION.

This section considers classification of the reals \mathbb{R} according to the exponent α in Eq 10.1. We will find a hierarchy, from the simplest rational fractions which are the most difficult ('unwilling to be approximated'), through quadratic and higher order irrationals to the transcendental numbers, which are the most 'willing'.

Eq (8) gives an exact expression for the difference Δ_k between two adjacent convergents. This was the starting point of a remarkable chain of thought which led Joseph Liouville to publish in 1844 a recipe for producing examples of transcendental numbers. Here is an outline.

Eq (10) states that

$$|\epsilon_k| = \left| C_F - \frac{p_k}{q_k} \right| < \frac{1}{a_{k+1}q_k^2} \leq \frac{1}{q_k^2}.$$

This is true of the convergents of *every* continued fraction, no matter what type of number it represents. We know that the convergents have the special 'best fit' property, and we found in §1.3, Table 1 some other fractions p/q which give good, efficient approximations to C_F . So how easy is it get good, efficient rational approximations to any given number? 'Efficient' here means that the denominator is in some sense small in relation to the accuracy of approximation. The larger the denominator, the more complicated and cumbersome the rational approximation, so q measures the investment of effort in obtaining accuracy of approximation. Are accurate, low denominator, rational approximations commonplace – ten a penny – or are they thin on the ground?

Take first the case of C_F being itself a rational number n/d . Its continued fraction representation is finite, so there can be only a finite number of convergents. Consider approximating n/d by another fraction p/q , not necessarily a convergent. If $n/d = p/q$, then $nq - pd = 0$, but otherwise

$$\left| \frac{n}{d} - \frac{p}{q} \right| = \left| \frac{nq - pd}{dq} \right| \geq \frac{1}{dq} > \frac{1}{q^2}$$

unless $d > q$. So the efficient approximations are restricted to values of q less than d . Taken altogether, the total number of fractions p/q which approximate to n/d well in the sense of Eq (10) is finite. Efficient approximations of rational numbers by other rationals are rare; most have errors larger than $1/q^2$. You could say that rationals with comparable denominators are quite well separated from each other on the real number line.

If θ is irrational, it has an infinite continued fraction and hence an infinite number of convergents which satisfy $|\epsilon| < 1/q^2$. But can we get even closer approximations efficiently? For instance, are there an infinity of approximations p/q with error less than $1/q^3$? Consider first the case of quadratic irrationals. These are numbers which involve square roots and have continued fractions with an infinitely recurring sequence. Since the convergents give the least error, it seems appropriate to search amongst them for cases of $\epsilon_k < 1/q_k^3$. Eq (11b) gives the lower bound on convergents $1/[(a_{k+1} + 2)q_k^2] < \epsilon_k$, so $\epsilon_k < 1/q_k^3$ would imply that $a_{k+1} + 2 \geq q_k$. Now while this might occur a couple of times within the first few convergents, it cannot be true for higher convergents since the q_k grow indefinitely and faster than $a_1 a_2 a_3 \dots a_k$ whilst the a_{k+1} remain bound at the largest value in the recursion sequence.

To take this line of thought further, what is the smallest value of θ , $0 < \theta < 1$ such that there is an infinity of rational approximations to a given quadratic irrational, each with error less than $1/q^{2+\theta}$? This is equivalent to asking for the smallest θ such that $a_{k+1} + 2 \geq q^\theta$ for all q . It's pretty obvious that for any non-zero θ , q^θ will exceed any given constant provided q is large enough – greater than $(a_{k+1} + 2)^{1/\theta}$. So θ must be 0.

To put the matter more rigorously, take first the limiting case of $\{1: \underline{1}\}$ which has the slowest rate of increase in q_k . For this the q_k form a Fibonacci series which eventually grows as the geometric series $\lambda^k/\sqrt{5}$ where $\lambda = \frac{1}{2}(1 + \sqrt{5})$. So we are asking for the lowest value of θ such that

$$\frac{\lambda^{\theta k}}{\sqrt{5}} < a_{k+1} + 2 = 3.$$

It is straightforward to show that, no matter how small is θ , q_k^θ will exceed 3 for every k greater than about $4/\theta$. Effectively, meeting the criterion requires that $\theta = 0$.

All other continued fraction have at least one a_k greater than 1, so converge faster. For these note from the recursion relation that $q_k \gg a_1 a_2 a_3 \dots a_k$. Suppose that $k = nL + \kappa$ where L is the length of the recursion sequence. If a_{max} is the largest of the a_k , then

$$q_k^\theta > (a_1 a_2 \dots a_l)^{n\theta} \cdot (a_1 a_2 \dots a_\kappa)^\theta \geq a_{max}^{n\theta}.$$

This will be greater than $a_{max} + 2$ if $n\theta > \ln(a_{max} + 2)/\ln(a_{max})$. The right side has its highest value of 2 for $a_{max} = 2$. We must conclude that for any convergent of any recurring continued fraction except $\{1: \underline{1}\}$, for all $n > 2/\theta$, $k > 2L/\theta + \kappa$, the error exceeds $1/q_k^{2+\theta}$. As for $\{1: \underline{1}\}$, effectively $\theta = 0$. Remembering that no rational can approximate to a given irrational better than its convergents, this means that there is only a finite, though possibly large, number of rational approximations p/q with $\epsilon < 1/q^2$. Moreover, these cannot have high values of q .

This concept is sometimes expressed by saying that quadratic irrationals are ‘infinitely approximable to order 2’, or have ‘approximation measure 2’ or ‘approximation exponent 2’. The 2 here denotes the highest value which the index of q can take and there still be an infinity of rational approximations with $\epsilon < 1/q^2$. The higher the index, the better the approximation you get for a given maximum value of q .

We are now in a position to state Liouville’s theorem. He extended the concept of finding the lowest index μ such that a given irrational \mathcal{I} of higher degree¹⁶ D than 2 would have an infinity of rational approximations satisfying

$$\left| \mathcal{I} - \frac{p}{q} \right| < \frac{K}{q^\mu} \quad (12)$$

where K is a positive constant depending on \mathcal{I} but independent of both p and q . Liouville’s theorem states that $\mu \leq D$.

The proof makes use of the fact that the minimal polynomial $P(x)$ of \mathcal{I} is a smooth continuous curve whose value and derivative over any finite interval are bounded. It also uses the fact that $P(\mathcal{I}) = 0$. Refer to Figure 24. It shows a graph representing $P(x)$ and two convergents to \mathcal{I} ; these necessarily lie either side of \mathcal{I} . We focus on approximation of \mathcal{I} by C_k and therefore on the distance $|\mathcal{I} - C_k|$.

¹⁶By ‘degree’ of \mathcal{I} is meant that \mathcal{I} is a zero of a polynomial with integer coefficients $P(x)$ of degree (highest power) D , and \mathcal{I} does not satisfy any polynomial equation of lower degree. $P(x)$ is called the minimal polynomial of \mathcal{I} .

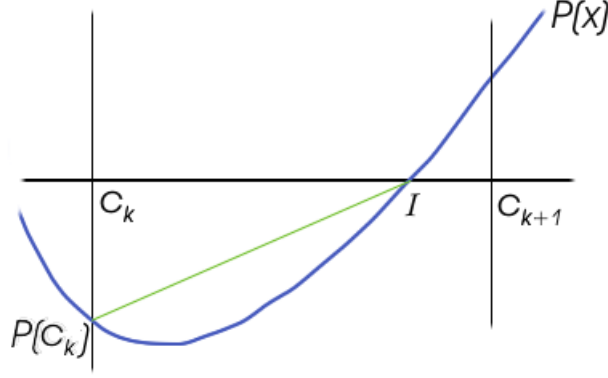


Figure 24: Illustrating the proof of Liouville's rational approximation theorem.

Suppose that $P(x) = J_D x^D + J_{D-1} x^{D-1} + \dots + J_0$ where the J_k are fixed integers. Since $C_k \neq \mathcal{I}$, $P(C_k) \neq 0$. In fact

$$|P(C_k)| = \frac{|J_D p_k^D + J_{D-1} p_k^{D-1} q_k + J_{D-2} p_k^{D-2} q_k^2 + \dots + J_0 q_k^D|}{q_k^D} \geq \frac{1}{q_k^D}.$$

Since $P(\mathcal{I}) = 0$,

$$|P(\mathcal{I}) - P(C_k)| = |P(C_k)| \geq \frac{1}{q_k^D}. \quad (13)$$

The Mean Value Theorem of calculus states that there is at least one point between C_k and \mathcal{I} where the gradient equals the gradient of the secant line (green in Figure 8) from $P(C_k)$ to $P(\mathcal{I})$. Since $P(x)$ is smooth and bounded at every point this gradient is less than some upper bound value $M_q > 0$:

$$\left| \frac{P(\mathcal{I}) - P(C_k)}{\mathcal{I} - C_k} \right| < M_q.$$

The subscript acknowledges that this limit is over the interval $[C_k, \mathcal{I}]$ and so does not satisfy the requirement of being independent of q . However, this obstacle is readily removed by taking a wider interval which includes C_k , such as $[\mathcal{I} - 1, \mathcal{I} + 1]$. Let the maximum gradient of $P(x)$ over this be M – it too is guaranteed to be finite and not less than the gradient of the secant in Figure 8.

Putting this together with Eq (13),

$$M|\mathcal{I} - C_k| > |P(\mathcal{I}) - P(C_k)| > \frac{1}{q_k^D} \quad \text{so}$$

$$|\mathcal{I} - C_k| > \frac{1}{M q_k^D}. \quad (14)$$

Liouville's theorem means that errors larger than $1/(M q_k^D)$ are the norm. An algebraic irrational \mathcal{I} of degree D therefore has only a finite number of approximations with exponent greater than D . One might consider these as a few 'lucky hits' on \mathcal{I} .

I have proved the theorem here for p_k/q_k being a convergent of \mathcal{I} and therefore a better approximation than any other fraction with equal or smaller denominator. With any other rational p/q the error will be larger, so Liouville's theorem is particularly true for non-convergents. In

summary, we have established these bounds on the error in approximating an algebraic number \mathcal{I} of degree D by almost all rationals p/q :

$$\frac{1}{Mq^D} < \left| \mathcal{I} - \frac{p}{q} \right| < \frac{1}{q^2}. \quad (14)$$

The upper bound applies to all rationals, and the lower bound to all but a finite number.

There is quite a lot to the interpretation of Eqs (12) and (14). They are consistent with the algebraic numbers being thinly spaced on the real number line, like the rationals themselves. (Hardy and Wright, p 160, explain how the algebraic numbers can be put in 1-1 correspondence with the integers.) For a fixed denominator q , as D increases, $1/q^D$ gets rapidly smaller, meaning that p/q can be a closer approximation to \mathcal{I} . Put the other way round, for any rational p/q there are numbers \mathcal{I} very close to p/q – closer than any algebraic number of low degree. Roughly speaking, algebraic numbers of high degree are more plentiful and hence more closely spaced on the real number line than those of low degree. For instance, to every quadratic expression one can prefix J_3x^3 for an infinite number of integers J , so there are in a sense ‘more’ cubic irrationals than quadratic ones. Taking the limit of degree D tending to infinity, we can picture trans-algebraic irrationals – the transcendental numbers. For these the gradient in Figure 8 can be infinite at arbitrary points, the constraint on the error $|\mathcal{I} - C_k|$ tends to zero, and so there must be transcendental numbers infinitesimally close to any and every rational p/q .

Liouville recognised that it is possible to manufacture numbers which have an infinite number of close rational approximations simply by ensuring that the error at every convergent is sufficiently small compared with q_k . This can be done by ensuring there is very little change from C_k to C_{k+1} , which in turn requires that the partial quotients increase by large amounts from a_k to a_{k+1} , without bound. Much work has been done since Liouville’s time to improve of the bounds in Eq (14). In 1908 Axel Thue made significant progress and in 1958 Klaus Roth won the Fields Medal for studies which concluded that the approximation index μ of all irrational algebraic numbers (not just quadratic ones) is exactly 2, and that for all transcendental numbers $\mu \geq 2$. For e , $\mu = 2$ exactly, the lowest value possible for a transcendental.

We can now use Roth’s criterion to manufacture examples of numbers which must be transcendental. Since for any convergent in the continued fraction of I ,

$$\left| \mathcal{I} - \frac{p_k}{q_k} \right| < \frac{1}{a_{k+1}q_k^2}$$

and since Roth’s criterion is that the error decreases as $1/q_k^{2+\theta}$, we can make I transcendental by making each a_{k+1} increase as a power of the last denominator, q_k . In the first numerical example, take $\theta = 1$, corresponding to $a_{k+1} = q_k$. Take the integer part, a_0 , to be 1. We thus build the fraction $\{1 : 1, 1, 2, 5, 27, 734, 538783, \dots\} = 1.592643049\dots$ whose convergents are respectively

$$\frac{1}{1}, \frac{2}{1}, \frac{3}{2}, \frac{8}{5}, \frac{43}{27}, \frac{1169}{734}, \frac{858089}{538783}, \text{ etc.}$$

In the second example, take $\theta = 1/3$ and make a_{k+1} equal to the nearest integer to $q_k^{1/3}$. For example $\sqrt[3]{3386931} = 150.1765$. If a_0 is 1, the fraction is

$$\{1 : 1, 1, 1, 1, 2, 2, 3, 5, 8, 17, 43, 150, 789, \dots\} = 1.613191116087\dots$$

Its convergents are

$$\frac{1}{1}, \frac{2}{1}, \frac{3}{2}, \frac{5}{3}, \frac{8}{5}, \frac{21}{13}, \frac{50}{31}, \frac{171}{106}, \frac{905}{561}, \frac{7411}{4594}, \frac{126892}{78659}, \frac{5463767}{3386931}, \frac{819691942}{508118309}.$$

Though the a_k increase slowly to start, they do increase rapidly later. There is an infinity of transcendental numbers which can be determined by this method, corresponding to a_{k+1} being different powers of q_k . Indeed, merely increasing one or two of the a_k by 1 will create a new transcendental number from either of the examples above ; for example $\{ 1 : 1, 1, 1, 2, 2, 2, 3, 5, 8, 17, 44, 151, 789, \dots \}$.

One should note that the a_{k+1} must increase rapidly with respect to q_k , and not just rapidly in some more general sense. For example, I had initially guessed that $\mathcal{I} = \{0 : 1^1, 2^2, 3^3, \dots k^k \dots\}$ would be transcendental, but find that I cannot show that it satisfies Roth's criterion, which is that $a_{k+1} \geq q_k^\theta$ for some $\theta > 0$. From the recursion relation $q_k > a_k a_{k-1} \dots a_2 a_1 a_0$, so for \mathcal{I} to be transcendental a necessary, though not sufficient, criterion is that

$$a_{k+1} = (k+1)^{(k+1)} \geq \left[k^k (k-1)^{(k-1)} (k-2)^{(k-2)} \dots 3^3 2^2 \right]^\theta \text{ meaning that}$$

$$\frac{(k+1) \log(k+1)}{k \log k + (k-1) \log(k-1) + (k-2) \log(k-2) + \dots + 3 \log 3 + 2 \log 2} \geq \theta.$$

But the left side tends to zero, slowly, as $k \rightarrow \infty$, forcing θ to be zero. The number \mathcal{I} might yet be transcendental, but it is not so proven by this analysis.

14 The cardinality of \mathbb{Q} , \mathbb{R} and \mathbb{R}^n

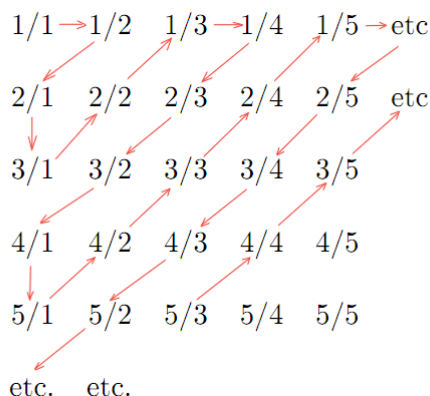
This section is something of an interesting asided. It describes further the history of how mathematicians in the nineteenth century came to understand more fully the types and properties of the real numbers. In the decades after Liouville had shown the existence of transcendentals, mathematicians in France and Germany were striving to place analysis, including the differential and integral calculus, on a rigorous footing. Central to this was finding a satisfactory definition of continuity. There was an awareness that although the rational numbers, \mathbb{Q} , were ‘dense’, meaning that between any two rationals there is at least one other rational¹⁷, they are not continuous. Continued fractions provided clear evidence of this, clearer than did decimals. This is because the continued fraction of a rational has strictly finite length, whereas its decimal representation has an infinity of recurring digits after the decimal point, particularly if, say $4 \cdot 63000\dots$ is written as $4 \cdot 629999\dots$.

In the 1870s two mathematicians in Germany, Georg Cantor and Richard Dedekind, were corresponding with each other on the concept of continuity and the intuitive belief that there are somehow more real irrational numbers than rational, and certainly more than the counting numbers \mathbb{N} . The \mathbb{N} are clearly discrete, $1, 2, 3, 4, \dots$, whereas the reals \mathbb{R} by definition continuously fill the number line.

Georg Cantor was a Russian-German mathematician who almost single handedly pioneered set theory. His classic proofs on cardinality were published in a series of papers between 1874 and 1895. He was analysing the cardinality of what we now call sets – that is, the number of elements they contain. Cantor used the concept of one-to-one correspondence to compare the size or ‘power’ of two sets. Where sets have an infinity of elements, such as the integers \mathbb{Z} , they are counted by being placed in a one-to-one correspondence with the natural number \mathbb{N} , the counting numbers:

$$\begin{array}{l|cccccccc} z \in \mathbb{Z} & -1 & 1 & -2 & 2 & -3 & 3 & -4 & 4 & \text{etc.} \\ n \in \mathbb{N} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & \text{etc.} \end{array}$$

Using his now famous diagonal counting scheme, Cantor also proved how the rationals \mathbb{Q} can be placed in one-to-one correspondence with \mathbb{N} , as shown below. Just follow the arrows as they weave diagonally through the array of rationals, counting as you go:



¹⁷The Farey series of §8.2 is a model for this.

This is all very counter-intuitive. The fractions are ‘dense’ on the real number line, yet here is proof that the ‘size’ of the set \mathbb{Q} is the same as that of \mathbb{N} . Cantor had to struggle to get his strange ideas accepted, facing open hostility from Kronecker is particular.

Equally surprising was Cantor’s placing of the algebraic numbers in one-to-one correspondence with \mathbb{N} (published in 1874). To do this he had to place the algebraic numbers in a series of ascending order. Algebraic numbers, sometimes denoted \mathbb{A} , are roots of polynomials with integer coefficients. Each has an irreducible polynomial of least degree, the minimal polynomial, to which it is a root. Cantor used an integer to measure the ‘altitude (höhe)’ of each minimal polynomial as the way to order the algebraic numbers¹⁸. His definition of altitude, \mathcal{L} , is the sum of the magnitudes of the coefficients, plus the degree, minus 1. There can be only a finite number of minimal polynomials with any given altitude value, depending on how \mathcal{L} can be partitioned into smaller integers, and the permutations of these integers as coefficients¹⁹. Thus, all minimal polynomials of fixed altitude \mathcal{L} can be ordered, and these sequences in turn ordered by ascending \mathcal{L} .

By 1873 Cantor had proved that the cardinality of the reals \mathbb{R} exceeds that of \mathbb{N} , and is a different and larger order of infinity. For this he needed a unique representation of any real number, r . He chose a decimal representation, subject to the rule that any number which would end in an infinite series of zeros, ‘0’, be replaced by an infinite series of ‘9’. For example $4 \cdot 60000 \dots$ is written as $4 \cdot 59999 \dots$. The proof runs as follows. Subject to the above rule, any real number r_1 has a unique representation as an infinite decimal $d_{1,0} \cdot d_{1,1} d_{1,2} d_{1,3} \dots$ and, conversely, each such sequence of digits represents one and only one real number. Suppose, for contradiction, that there is a one-to-one correspondence between \mathbb{R} and \mathbb{N} . Then every real can be indexed uniquely by $n \in \mathbb{N}$, and the entire set \mathbb{R} written as a table of decimals starting

$$\begin{array}{lcl}
 r_1 & = & d_{1,0} \cdot d_{1,1} d_{1,2} d_{1,3} d_{1,4} d_{1,5} d_{1,6} \dots \\
 r_2 & = & d_{2,0} \cdot d_{2,1} d_{2,2} d_{2,3} d_{2,4} d_{2,5} d_{2,6} \dots \\
 r_3 & = & d_{3,0} \cdot d_{3,1} d_{3,2} d_{3,3} d_{3,4} d_{3,5} d_{3,6} \dots \\
 r_4 & = & d_{4,0} \cdot d_{4,1} d_{4,2} d_{4,3} d_{4,4} d_{4,5} d_{4,6} \dots \\
 r_5 & = & d_{5,0} \cdot d_{5,1} d_{5,2} d_{5,3} d_{5,4} d_{5,5} d_{5,6} \dots \\
 \text{etc} & & \text{etc}
 \end{array}$$

Bear in mind that *every* real number is supposed to be listed in this table. The $d_{i,j}$ range from 0 to 9, though there is never a tail of zeros. Now invent a new number s with decimal $b_0 \cdot b_1 b_2 b_3 \dots$ in which $b_0 \neq d_{0,0}$, $b_1 \neq d_{1,1}$, $b_2 \neq d_{2,2}$, etc. The number of reals is supposedly countably infinite but so also is the number of decimal places in each. Therefore every real r_j has a decimal place at which it will differ from s . In this way s differs from every r_j in at least one decimal digit, the j^{th} , and so cannot itself be one of the r_j . So $s \notin \mathbb{R}$, in contradiction to the assumption that every real is contained in the above table. In other words, \mathbb{R} is not countable by \mathbb{N} .

Cantor surprised even himself by next proving that the cardinality of \mathbb{R} , the real number line, equals that of \mathbb{R}^n , the Euclidean space of n dimensions. In a ground-breaking paper of 1878 he gave his proof using infinite continued fractions (not decimals). He had initially formulated a proof using infinite decimals, as above, but Dedekind had pointed out the problem with the

¹⁸This is often translated as ‘height’, but height is used nowadays to mean the largest absolute value of the coefficients.

¹⁹As an example, consider the degree 2 polynomial $ax^2 + bx + c$, which could be minimal. If the sum of magnitudes of the coefficients is fixed at 4, only 14 minimal polynomials can be formed with $(a, b, c) = (3, 0, \pm 1), (1, 0, \pm 3), (0, 3, \pm 1), (0, 1, \pm 3), (2, \pm 1, -1), (1, \pm 2, -1), (1, \pm 1, 2)$. Other degree 2 polynomials factorise.

ambiguity of $4 \cdot 0000 \dots = 3 \cdot 9999 \dots$ etc. which would cause double counting and so prevent one-to-one correspondence. Cantor quickly revised his proof to using continued fractions, as outlined below. It is noteworthy that some modern accounts of this present the proof using infinite decimal representations of the reals, as Cantor himself had done initially. Cases include Birkhoff & MacLane 'A Survey of Modern Algebra' page 363, and Seymour Lipschutz in 'Set Theory, Schaum's Outline Series' page 158.

Consider in the first case a mapping from the open unit interval $(0, 1)$ in \mathbb{R} to the open unit square in \mathbb{R}^2 , (x, y) , $0 < x < 1$, $0 < y < 1$. To be precise, the mapping carries (x, y) with *irrational* numbers as co-ordinates to an *irrational* point r in $(0, 1)$. Let $x \in (0, 1)$ have the continued fraction $\{0 : x_1, x_2, x_3, \dots\}$ and $y = \{0 : y_1, y_2, y_3, \dots\}$, so that together these uniquely specify one point in the unit square. Form from x and y

$$r = \{0 : r_1, r_2, r_3, \dots\} = \{0 : x_1, y_1, x_2, y_2, x_3, y_3, \dots\}.$$

In this way (x, y) is mapped to a unique point r , and the map has a unique inverse. Taking all such points, every point with irrational co-ordinates over the unit square is placed in one-to-one correspondence with an irrational point in $(0, 1)$. This proves that the two sets have the same cardinality. By extension of the same argument, irrational points (x, y, z) within the unit cube are placed in one-to-one correspondence with the unit interval by the continued fraction map

$$(x, y, z) \rightarrow \{0 : x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3, \dots\}.$$

In this way, using the unique representation property of continued fractions, Cantor proved that the *irrational* numbers within \mathbb{R} have the same cardinality as the irrational numbers in \mathbb{R}^n for all dimensions n – another totally counter-intuitive fact.

I italicise *irrational* because the argument does not carry over to rationals. To see this, suppose that the rationals, with their finite continued fractions, were included in Cantor's argument. Consider the types of co-ordinates which would then be mapped by $\{0 : x_1, y_1, x_2, y_2, x_3, y_3, \dots\} \rightarrow \{0 : r_1, r_2, r_3, \dots\}$. These following points show that the map $(x, y) \rightarrow r$ preserves in part the types of number:

1. If both x and y are fractions, with a finite number of partial quotients, the continued fraction for r must also be of finite length, making r rational too.
2. Suppose x is rational with F partial quotients and y is irrational. Then the first $2F$ partial quotients of r will be an interweaving of the first F from each of x and y , but thereafter r will have the same tail as y . From the discussion below in §8.3 Part III, y and r are 'equivalent' numbers, related via a matrix transformation with determinant ± 1 .
3. If both x and y are quadratic surds with recursion periods L_x, L_y , r will be a quadratic surd with recursion length $\leq 2L_x L_y$.
4. If both, or only one of x, y is a non-quadratic irrational, so must be r .
5. The above points hold if the continued fraction expansions of x and y have no digits in common. Where digits are shared, there may not be one-to-one correspondence.

Regarding point 5), it is clear that if $x = \{0 : a, a\}$ and $y = \{0 : a, a, a\}$, then $(x, y) \rightarrow \{0 : a, a, a, a, a\}$. But several other x, y pairs also map to this number: $\{0 : a\}$ and $\{0 : a, a, a, a\}$ for instance. In

particular every pair with $x = 1/G = \{0 : \underline{1}\}$ and y any convergent of $1/G$ will always map to $1/G = \frac{1}{2}(\sqrt{5}-1)$. There is a countable infinity of convergents of $1/G$, each of the form $\{0 ; 1, 1, \dots, 1, 1\}$, so this mapping is far from being bijective.

Cantor realised there was a problem with rationals, so produced a supplementary argument to show that the set of irrationals on the open interval $(0, 1)$ could be put in one-to-one correspondence with all the reals, both irrational and rational, over the closed interval $[0, 1]$. He was the first to admit that this marred the simple elegance of the proof, but needs must. His argument drew upon the proof he had given much earlier in his career, that the rationals are denumerable. It runs as follows. Label the rational fractions f_1, f_2, \dots in order. From the irrationals pick out a sequence e_1, e_2, \dots indexed by $j \in \mathbb{N}$. Cantor's choice was $\sqrt{2}/2^j$ where n is a positive integer, but $\pi/(j+3)$ or others would serve equally well. Call H the set of all irrationals in $(0, 1)$ except e_1, e_2, \dots .

The set of all irrationals in $(0, 1)$ is $H \cup \{e_1, e_2, \dots\}$

whilst the set of all reals on $[0, 1]$ is $H \cup \{e_1, e_2, \dots\} \cup \{f_1, f_2, \dots\}$.

Now the trick. Separate the e_j into two subsets having odd and even index j , and pair the odd with the whole set $\{e_j\}$, and pair the even with the set $\{f_j\}$:

e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	etc.
e_1	f_1	e_2	f_2	e_3	f_3	e_4	f_4	e_5	etc.

Now the set of all reals on $[0, 1]$ is $H \cup \{e_1, e_3, \dots\} \cup \{e_2, e_4, \dots\}$.

which is in one-to-one correspondence with the irrational only. Including the rationals with the irrationals has made no discernable difference to the cardinality of \mathbb{R} . Today we would say that \mathbb{Q} has 'measure zero' in \mathbb{R} .

Cantor was so surprised by his proof that \mathbb{R} and \mathbb{R}^n have the same cardinality that he remarked "I see it, but I don't believe it". He thought it challenged the very concept of dimension, arguing that, because 2 co-ordinates could be replaced by 1, a 2-D surface could be replaced by a line segment, and so on for spaces spanned by a higher number of co-ordinates. However, Dedekind talked him out of this extreme view.

I do not know why Cantor's was so convinced by Dedekind's objection to a decimal-based argument, when such authoritative modern textbooks as Birkhoff and MacLane give it. Cantor's own proof using continued fractions seems forgotten. Perhaps it reflects how continued fractions, like slide rules and log tables, have disappeared from the tool kit of today's mathematicians.

PART IV is not completed.

Part IV

Patterns and Statistics in Partial Quotients

15 Patterns in a_k for finite continued fractions

This section looks at some patterns and statistical properties of continued fractions of rational numbers, focusing on the partial quotients a_k . Related statistical aspects of generic infinite continued fractions are deferred until §??

15.1 Observed patterns in partial quotients

To gain an overview, I have evaluated the continued fraction representation $\{0 : a_1, a_2, \dots, a_F\}$ of all fractions p/q for $2 \leq q \leq 1100$ and $1 \leq p \leq q - 1$, plus selected other rationals. Interesting patterns are thereby revealed

PART IV is not completed.