Essays on the arithmetic of quadratic fields

Bill Casselman
University of British Columbia
cass@math.ubc.ca

Approximating irrational numbers by rational ones

I intend this essay to be a brief introduction to continued fractions. There is a huge literature on this topic, but my approach will be more geometric than most. This has the potential to make many points more transparent, although whether my current version does so is moot.

There are several rather different reasons why continued fractions are interesting. One is that they produce very good rational approximations to real numbers. In practice this doesn't seem to be all that useful, but continued fractions have been used to prove certain numbers (for example e) to be transcendental. (For more about this, look at the `mathoverflow` discussion mentioned in the list of references.). Another important application, and my immediate motivation, is to the arithmetic of quadratic extensions of $\mathbb{Q}$.

Connections with other branches of mathematics are described briefly in the Wikipedia entry on continued fractions. In addition to *arithmetic* continued fractions, there are *analytic* continued fractions, in which continued fractions involving functions are used to approximate quite general functions. I'll say nothing about these, beyond the remark that one place to see at least some information on this topic is the Wikipedia entry on Padé approximants. There are also applications to combinatorics.

Contents

1. Introduction

Suppose $\lambda > 0$ to be irrational. Given a positive integer q, the number λ lies in one of the half-open intervals $[p/q, (p+1)/q)$, in which case

$$\left|\lambda - \frac{p}{q}\right| < \frac{1}{q} .$$

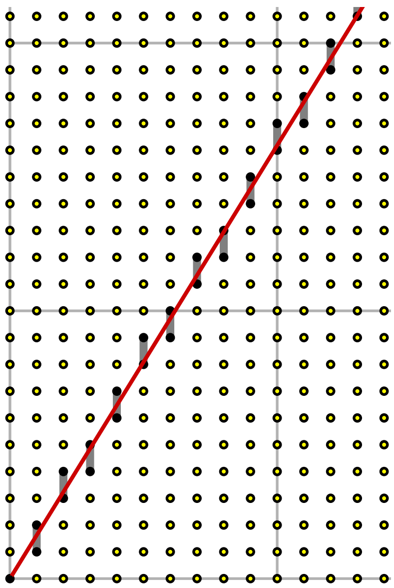
The inequality above is equivalent to

$$|p - q\lambda| < 1 .$$

Given q, finding p is simple since

$$\lfloor q\lambda \rfloor < q\lambda < \lfloor q\lambda \rfloor + 1 .$$

This can be pictured—the claim is just that the line $y = \lambda x$ meets the line $x = q$ at a distance less than 1 from the point (q, p). The figure below illustrates how things look when λ is the golden ratio $\tau = (1 + \sqrt{5})/2 \sim 1.618$.



How lattice points approximate $\boldsymbol{y = \tau x}$

This is the content of decimal approximations: $e - 2718/1000 < 1/1000$. Of course we can do slightly better without extra work, getting a gap of $1/2q$ rather than $1/q$. This is not significant. But the figure also shows that the way in which lattice points approach the line $y = \tau\, x$ is rather erratic. There are evidently some which are very close to the line, meaning that some fractions approximate τ very closely. The following well known result tells us that we can do arbitrarily well:

1.1. Proposition. *There exist points of the lattice arbitrarily close to the line $y = \lambda x$, on both top and bottom.*

This is equivalent to the superficially more general:

1.2. Proposition. *If λ is irrational, the points $n\lambda$ modulo 1 are dense in* $[0, 1]$.

Proof. I'll first show that for any $N > 0$ there exists m, n such that $|n\lambda - m| < 1/N$. Plot the $N + 1$ numbers $n\lambda - \lfloor n\lambda \rfloor$ for $1 \le n \le N + 1$. Because there are N intervals $[i/N, (i + 1)/N$ in $[0, 1]$ and $N + 1$ of these numbers, at least two of the numbers must lie in the same interval, say for $n_1\lambda$ and $n_2\lambda$ with $n_2 > n_1$. But then there exists some integer m such that

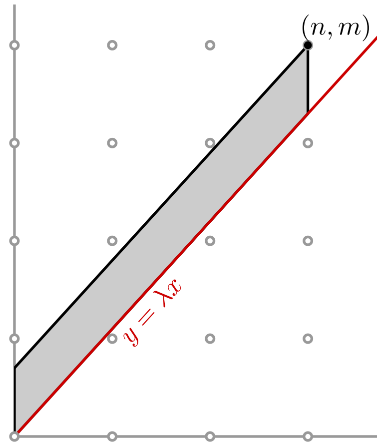
$$|(n_2\lambda - n_1\lambda) - m| = |(n_2 - n_1)\lambda - m| < 1/N .$$

If $n\lambda = m + \varepsilon$ with $|\varepsilon| < 1/N$ then each interval of width $1/N$ will contain a number of the sequence $kn\lambda$ modulo 1.

There is a very simple if inefficient procedure for finding good approximations, just scanning horizontally from one approximation until you find a closer one. However, we can do much better.

2. Continued ...

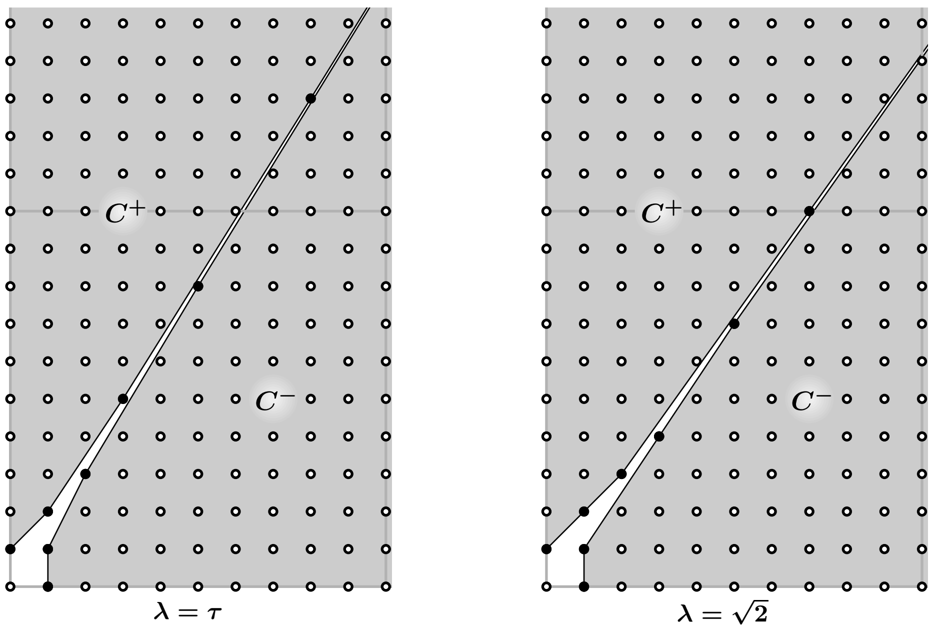
It is natural to ask, what are the best rational approximations to any given real number λ? In a precise formulation, I call (m, n) with $n > m\lambda$ **optimal** if (m, n) is closer to the line $y = \lambda x$ than any other point (q, p) such that $p > \lambda q$ and $q < n$. Similarly for a point with $n < m\lambda$. In the following figure, the point (n, m) is optimal.



How can one find optimal points? Continued fractions give a good answer to this question. I'll present here a graphical version of the method due implicitly to the nineteenth century English mathematician Henry J. Smith (in a casual observation at the end of [Smith:1876]) and explicitly to Felix Klein.

Let λ^+ be the set of all lattice points in the positive quadrant above the line $y = \lambda x$, and λ^- be that of all lattice points in the same quadrant below it. Let $C^\pm$ be the corresponding convex hulls. The space between C^+ and C^- contains no lattice points, since λ is irrational. Let ∂^+ be the bottom boundary of C^+, ∂^- the upper boundary of C^-. Let $\mathfrak{d}^+$ be the set of lattice points on ∂^+, $\mathfrak{d}^-$ those on ∂^-

The following figures illustrate these sets for $\lambda = \tau$ and $\sqrt{2}$.



2.1. Lemma. *If $P = (q, p)$ is a point of ∂^+, then in the region $x \geq q$ the line $y = \lambda x$ lies below the line through P parallel to it. Similarly for a point of ∂^-.*

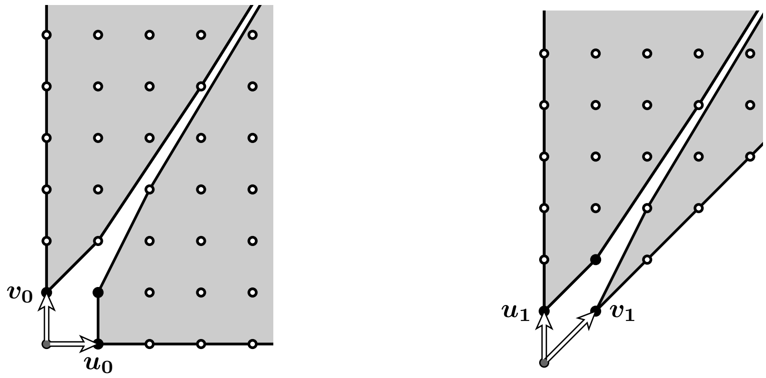
Proof. Both C^+ and C^- are convex. Proposition 1.1 implies that each of them approaches arbitrarily closely to the line $y = \lambda x$.

The points of the sets $\mathfrak{d}^\pm$ will certainly give good rational approximations to λ, and it turns out that there is a very efficient way to find them. The starting point is this elementary observation:

2.2. Lemma. *Set $\ell = \lfloor \lambda \rfloor$. Then*

(a) $(0, 1)$ *is a vertex of* C^+*;*
(b) the points $(1, i)$ *for* $0 \leq i \leq \ell$ *lie in* $\mathfrak{d}^-$*, and the endpoint* $(1, \ell)$ *is a vertex of* C^-*;*
(c) the points listed above are all the points of C^- *in the region* $x \leq 1$*.*

The coordinates here are, of course, in the standard rectangular system.



But now I change the basis of the lattice $\mathbb{Z}^2$. To make notation simpler, set $\lambda_0 = \lambda$ and $\ell_0 = \lfloor \lambda \rfloor$. The old basis is $u_0 = (1, 0)$ and $v_0 = (0, 1)$, while the new one is

$$\begin{aligned} u_1 &= v_0 = (0, 1) \\ v_1 &= u_0 + \ell_0 v_0 = (1, \ell_0) \, . \end{aligned}$$

In other words, I shear the original basis and reverse orientation as well. With this new basis, we have a new coordinate system (x_1, y_1). The associated change of coordinates is

$$\begin{aligned} x_0 &= y_1 \\ y_0 &= x_1 + \ell_0 y_1 \, . \end{aligned}$$

We can substitute and solve for y_1.

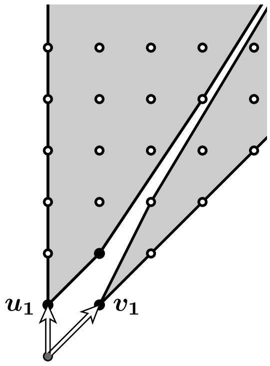
- *The line $y_0 = \lambda_0 x_0$ is now the line*

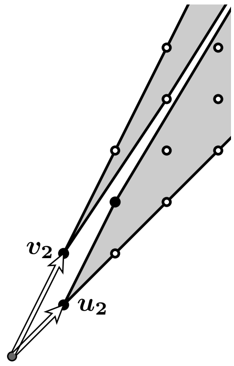
$$y_1 = \lambda_1 x_1 \, .$$

if $\lambda_1 = 1/(\lambda_0 - \ell_0)$.

The next basic observation is:

> *The points of $\mathfrak{d}^\pm$ in the new coordinate system are the same as those of the old, except that we have lost all the points on the line $x = 1$ in the old system, other than the vertex v_1.*





In other words, we are in essentially the same situation as when we started out.

This leads to an infinite inductive process. We may continue on forever to find new boundary lattice points and new bases, as well as edges connecting them to one of the boundary points already found. Sooner or later every boundary lattice point will arise. The extremal points among these are vertices of $C^{\pm}$. Expressing the points we find in the original coordinate system, we find the approximations we are looking for, as in the classical form of continued fractions.

Restricting ourselves to vertices, we get in this way a sequence of bases (u_n, v_n):

$$\begin{aligned}
u_0 &= (1, 0) \\
v_0 &= (0, 1) \\
\lambda_0 &= \lambda \\
\ell_0 &= \lfloor \lambda_0 \rfloor \\
\\
u_1 &= (0, 1) \\
v_1 &= (1, \ell_0) \\
\lambda_1 &= 1/(\lambda_0 - \ell_0) \\
\ell_1 &= \lfloor \lambda_1 \rfloor \\
&\ \dots \\
u_n &= v_{n-1} \\
v_n &= u_{n-1} + \ell_{n-1} v_{n-1} \\
\lambda_n &= 1/(\lambda_{n-1} - \ell_{n-1}) \\
\ell_n &= \lfloor \lambda_n \rfloor \,.
\end{aligned}$$

Note that $u_n + \ell_n v_n = v_{n+1} = u_{n+2}$.

As a consequence of this construction and Lemma 2.2:

2.3. Theorem. *Each set $\partial^{\pm}$ is the union of segments u, $u + v$ in which u and v make up a lattice basis, and one each in $\mathfrak{d}^{\pm}$.*

Define sequences (p_n), (q_n) by the specification $u_n = (q_{n-2}, p_{n-2})$. The recursive formula for the vectors u_n then give us initial conditions and recursive formulas. The basic step goes

(2.4)
$$\begin{aligned}
p_n &= p_{n-1}\ell_n + p_{n-2} \\
q_n &= q_{n-1}\ell_n + q_{n-2} \,.
\end{aligned}$$

One can make the process into a simple worksheet:

n	λ_n	ℓ_n	p_n	q_n
-2			0	1
-1			1	0
0	λ	$\ell_0 = \lfloor \lambda_0 \rfloor$	1	1
		$\dots$		
$n-2$			p_{n-2}	q_{n-2}
$n-1$	λ_{n-1}	ℓ_{n-1}	p_{n-1}	q_{n-1}
n	$\lambda_n = 1/(\lambda_{n-1} - \ell_{n-1})$	$\ell_n = \lfloor \lambda_n \rfloor$	$p_{n-1}\ell_n + p_{n-2}$	$q_{n-1}\ell_n + q_{n-2}$

For example with $\lambda = \sqrt{2}$:

-2			0	1
-1			1	0
0	$\sqrt{2}$	1	$\mathbf{1}$	1
1	$1+\sqrt{2}$	2	$\mathbf{3}$	2
2	$1+\sqrt{2}$	$\mathbf{2}$	$\mathbf{7 = 2 \cdot 3 + 1}$	5

I recall that the p_n/q_n are the successive approximations to the original λ.

The recursion formulas can be encapsulated in terms of matrices

$$\Gamma_n = [\, u_n \quad v_n \,]$$

whose columns are the vectors in the n-th basis:

$$\begin{aligned} \Gamma_0 &= [\, u_0 \quad v_0 \,] \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ \Gamma_1 &= [\, u_1 \quad v_1 \,] \\ &= \begin{bmatrix} 0 & 1 \\ 1 & \ell_0 \end{bmatrix} \\ &\dots \\ \Gamma_n &= [\, u_n \quad v_n \,] \\ &= \Gamma_{n-1} \begin{bmatrix} 0 & 1 \\ 1 & \ell_{n-1} \end{bmatrix} . \end{aligned}$$

Induction shows that $\det(\Gamma_n) = (-1)^n$. Since

$$\Gamma_n = \begin{bmatrix} q_{n-2} & q_{n-1} \\ p_{n-2} & p_{n-1} \end{bmatrix} ,$$

this leads to an agreeable fact:

2.5. Proposition. *The numbers p_n, q_n are relatively prime.*

Examples. (1) If λ is the root of a quadratic polynomial with rational coefficients, its continued fraction is eventually repetitive. For example, say $\lambda = \lambda_0 = \tau$. Then

$$\begin{aligned} \ell_0 &= \left\lfloor \frac{1+\sqrt{5}}{2} \right\rfloor \\ &= 1 \\ \lambda_1 &= \frac{1}{\lambda_0 - 1} \\ &= \frac{1}{(-1+\sqrt{5}/2)} \\ &= \frac{2(-1-\sqrt{5})}{1-5} \\ &= \tau . \end{aligned}$$

At this point things repeat, and in this case $\ell_n = 1$ for every n. Similar periodicity occurs for every quadratic irrational number. The earliest systematic reference to this phenomenon that I know of is [Euler:1744].

(2) Other examples for which the behaviour of the sequence (ℓ_n) is regular or even predictable are rare. [Euler:1744] finds the sequence of the ℓ_n for several numbers related to e. For e itself it is

$$2,\ 1,\ 2,\ 1,\ 1,\ 4,\ 1,\ 1,\ 6,\ 1,\ 1,\ 8,\ \ldots$$

This was a remarkable discovery, because ordinary floating point calculations to get a long and accurate expansion require high precision. Euler even managed to prove the expansion to be correct, by an even more remarkable analysis (in §§28 *ff.*) of solutions of Riccati's differential equation. (This is related to the subject of analytic continued fractions I mentioned in the Preface.)

In contrast, the continued fraction of π shows no comprehensible pattern.

3. ... fractions

The equation relating λ_n and λ_{n+1} can be reformulated as

(3.1)
$$\lambda_n = \ell_n + \frac{1}{\lambda_{n+1}}\,.$$

We then deduce by induction a sequence of equations:

(3.2)
$$\begin{aligned}\lambda &= \lambda_0\\ &= \ell_0 + \frac{1}{\lambda_1}\\ &= \ell_0 + \cfrac{1}{\ell_1 + \cfrac{1}{\lambda_2}}\\ &= \ell_0 + \cfrac{1}{\ell_1 + \cfrac{1}{\ell_2 + \cfrac{1}{\lambda_3}}}\\ &\ldots\end{aligned}$$

If p_n/q_n is the corresponding n-th rational approximation, then

$$\begin{aligned}p_0/q_0 &= \ell_0\\ p_1/q_1 &= \ell_0 + \frac{1}{\ell_1}\\ p_2/q_2 &= \ell_0 + \cfrac{1}{\ell_1 + \cfrac{1}{\ell_2}}\\ p_3/q_3 &= \ell_0 + \cfrac{1}{\ell_1 + \cfrac{1}{\ell_2 + \cfrac{1}{\ell_3}}}\\ &\ldots\end{aligned}$$

This explains the term 'continued fractions'. To avoid typesetting problems, I'll express the first, for example, as

$$\langle\!\langle \ell_0, \ell_1, \ell_2, \lambda_3 \rangle\!\rangle\,.$$

Another commonly used expression is

$$\ell_0 + \frac{1}{\ell_1 +}\,\frac{1}{\ell_2 +}\,\frac{1}{\lambda_3}\,.$$

I'll now be a bit more formal. If (x_i) is any finite array of real numbers, define $\langle\!\langle x_0, \ldots x_n \rangle\!\rangle$ by induction on n:

(3.3)
$$\begin{aligned}\langle\!\langle x_0 \rangle\!\rangle &= x_0\\ \langle\!\langle x_0, \ldots, x_n \rangle\!\rangle &= x_0 + \frac{1}{\langle\!\langle x_1, \ldots, x_n \rangle\!\rangle}\,.\end{aligned}$$

This recipe makes sense even if $x_i = \infty$, with the understanding that $1/\infty = 0$.

It is not difficult to verify a second inductive formula:

(3.4)
$$\langle\!\langle x_0, \ldots, x_{n-1}, x_n \rangle\!\rangle = \langle\!\langle x_0, \ldots, x_{n-1} + 1/x_n \rangle\!\rangle\,.$$

3.5. Lemma. *Suppose x_0 to be an integer and that $x_i \geq 1$ for $i \geq 0$. Then*

$$x_0 = \lfloor \langle\!\langle x_0, \ldots, x_n \rangle\!\rangle \rfloor$$

unless $n = 1$ and $x_1 = 1$. In the exceptional case

$$x_0 = \lfloor \langle\!\langle x_0, \ldots, x_n \rangle\!\rangle \rfloor - 1 \, .$$

Proof. Because of (3.4) , since $\langle\!\langle x_0, \ldots, x_n \rangle\!\rangle > 1$ if $n \geq 1$.

3.6. Proposition. *If $\lambda > 0$ is irrational then there is exactly one sequence (ℓ_n) such that*

$$\lambda = \langle\!\langle \ell_0, \ell_1, \ell_2, \ldots \rangle\!\rangle \, .$$

Proof. This follows immediately from the Lemma.

3.7. Proposition. *If λ is rational there are exactly two such expressions. Given $\varepsilon = \pm 1$ then there is exactly one finite array such that*

$$\lambda = \langle\!\langle \ell_0, \ldots, \ell_{n-1} \rangle\!\rangle$$

with $(-1)^n = \varepsilon$.

Proof. Suppose

$$\mathfrak{cf}(\lambda) = (\ell_0, \ldots, \ell_{n-1} \, .$$

Here, by construction, $\ell_{n-1} \geq 2$. Thus

$$\lambda = \langle\!\langle \ell_0, \ldots, \ell_{n-1} \rangle\!\rangle \, .$$

It follows from the Proposition that

$$\lambda = \langle\!\langle \ell_0, \ldots, \ell_{n-1} - 1, 1 \rangle\!\rangle$$

is the unique other expression for λ as a continued fraction. Exactly one of these will match ε in parity.

The number $\langle\!\langle x_0, \ldots, x_n \rangle\!\rangle$ can be calculated most conveniently by following the inductive recipe (2.4) . That is to say

$$\langle\!\langle x_0, \ldots, x_n \rangle\!\rangle = \frac{p_n}{q_n}$$

with the (p_n), (q_n) computed, as before, by induction:

$$\begin{aligned} p_0 &= x_0 \\ q_0 &= 1 \\ &\ldots \\ p_n &= p_{n-1} x_n + p_{n-2} \\ q_n &= q_{n-1} x_n + q_{n-2} \, . \end{aligned} \tag{3.8}$$

This process associates to every λ an infinite array $(\ell_n) = \mathfrak{cf}(\lambda)$. It is defined by an inductive definition of pairs:

$$\begin{aligned} \lambda_0 &= \lambda \\ \ell_0 &= \lfloor \lambda \rfloor \\ \lambda_n &= \frac{1}{\lambda_{n-1} - \ell_{n-1}} \\ \ell_n &= \lfloor \lambda_n \rfloor \, , \end{aligned}$$

with the convention that $1/0 = \infty$. The only case in which this will be invoked is when λ is a rational number and the continued fraction is finite.

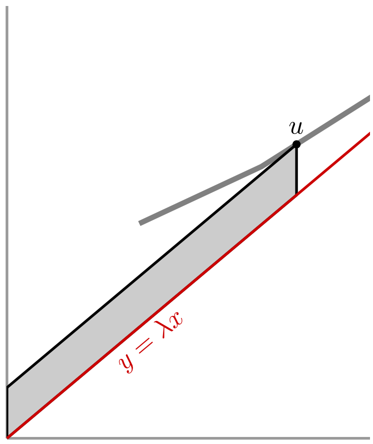
The continued fraction of λ may obviously be evaluated when λ is rational. We shall see in a later section that it may also be assigned a value as a limit otherwise. In all cases, the evaluation will be equal to λ.

4. Optimal approximations

One of the main results of this essay:

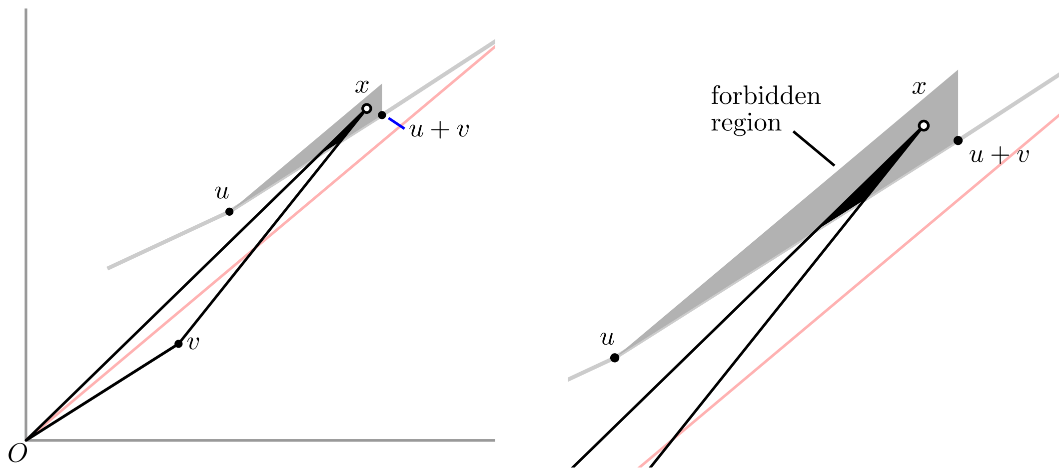
4.1. Theorem. *The optimal rational approximations to the line $y = \lambda x$ are the $\mathfrak{d}^{\pm}$.*

Proof. First of all, if u is a lattice point in $\mathfrak{d}^{+}$, the convexity of C^{+} implies that u is optimal.

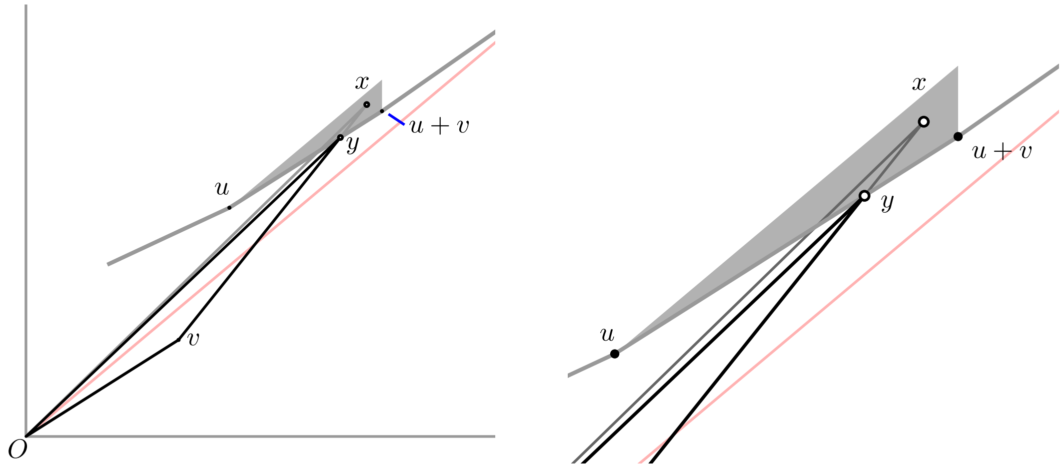


This is illustrated in the figure above.

Why are all the optimal points among the $\mathfrak{d}^{\pm}$? Working with both $\mathfrak{d}^{+}$ and $\mathfrak{d}^{-}$ is similar, so I'll look only at $\mathfrak{d}^{+}$. It is to be shown that any optimal point x lies on the bottom boundary of C^{+}. If not, suppose it lies to the right of u and the left of $u + v$, with u in $\mathfrak{d}^{+}$ and v in $\mathfrak{d}^{-}$. If it is not in $\mathfrak{d}^{+}$, it must lie in the region shaded lightly as below.



The argument depends on two well known facts about lattices, consequences of a theorem of Minkowski. Suppose u, v to be points in the lattice. (1) They form a basis if and only if the closed triangle Ouv contains no lattice points other then O, u, and v. (2) If form a basis, the area of the closed triangle Ouv is $1/2$.



The first criterion is satisfied for the points Oxv. By definition of C^+, any lattice point in this triangle would have to be in the region shaded black on the left. But any point in that region would be more optimal than x, a contradiction.

So Oxv is a basis of the lattice. But now I claim that the area of this triangle is more than $1/2$. This is because if w is any point on the segment from u to $u + v$, then the area of Owv is equal to $1/2$. But if y is chosen to be the intersection of the line xv with uv, then on the one hand the areas Oyv is $1/2$, while on the other it is strictly contained in Oxv.

5. Convergence

The formula (3.8) asserts that

$$\langle\!\langle x_0, \ldots, x_n \rangle\!\rangle = \frac{p_{n-1}x_n + p_{n-2}}{q_{n-1}x_n + q_{n-2}} \,.$$

Suppose $(x_0, \ldots, x_n)$ and y_n given. Then

$$\begin{aligned} \langle\!\langle x_0, \ldots, x_{n-1}, x_n \rangle\!\rangle - \langle\!\langle x_0, \ldots, x_{n-1}, x_n \rangle\!\rangle &= \frac{p_{n-1}x_n + p_{n-1}}{q_{n-1}x_n + q_{n-2}} - \frac{p_{n-1}y_n + p_{n-1}}{q_{n-1}y_n + q_{n-2}} \\ &= \frac{(p_{n-1}q_{n-2} - p_{n-2}q_{n-1})(x_n - y_n)}{(q_{n-1}x_n + q_{n-2})(q_{n-1}y_n + q_{n-2})} \\ &= \frac{\det(\Gamma_n)\cdot(x_n - y_n)}{(q_{n-1}x_n + q_{n-2})(q_{n-1}y_n + q_{n-2})} \,. \end{aligned} \tag{5.1}$$

This has two applications.

5.2. Proposition. *If $x_n \ge 1$ for all n then the sequence*

$$\langle\!\langle x_0, x_1, x_2, \ldots \rangle\!\rangle$$

converges.

Proof. I apply Cauchy's criterion. If $x_i \ge 1$ for all i then $q_n \ge n$ for all n, and $q_n q_{n+1} \ge n^2$. Hence

$$\sum \frac{1}{q_n q_{n+1}} < \infty$$

and the difference between any two terms in the sequence eventually differ by an arbitrarily small amount.

5.3. Proposition. *If $(\ell_i) = \mathfrak{cf}(\lambda)$ and $\mu = \langle\!\langle \ell_0, \ldots, \ell_n \rangle\!\rangle$ then*

$$|\lambda - \mu| \le \frac{1}{q_n^2} \,.$$

Proof. Since by (3.2)

$$\lambda = \langle\!\langle \ell_0, \ldots, \ell_{n-1}, \lambda_n \rangle\!\rangle$$

this follows directly from (5.1) , since $|\lambda_n - \ell_n| \le 1$.

It is worth noting that the convergence is actually fairly rapid. If $\lambda = \tau$, then $\ell_n = 1$. In this case the p_n and q_n are among the Fibonacci numbers f_n

$$1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34,\ \ldots$$

with

$$f_0 = 1, \quad f_1 = 1, \quad f_n = f_{n-1} + f_{n-2} \,.$$

This satisfies a difference equation of order two with constant coefficients. Since the roots of the indicial equation

$$x^2 - x - 1 = 0$$

are τ and $\overline{\tau}$, the solutions are linear combinations

$$\alpha\tau^n + \beta\overline{\tau}^n \,.$$

With the given initial conditions this gives us the solution with

$$\alpha = \frac{1 - \overline{\tau}}{\tau - \overline{\tau}}, \quad \beta = \overline{\alpha} \,.$$

Finally. we see that p_n and q_n both increase in magnitude roughly as powers of τ. So the convergence of the continued fraction is also geometric.

6. Rational numbers

I have so far assumed that λ is irrational, but that is unnecessary. In that case, if $\lambda = p/q$ the procedure given above amounts to a particular implementation of the Euclidean algorithm to findthe greatest common divisor of p, q. Everything in (2.4) works fine, say with $p > q > 0$ relatively prime, up intil the point when $\lambda_n = \ell_n$, because the next step would be to divide by 0. At that moment

$$\frac{p}{q} = \frac{p_n}{q_n} = \frac{p_{n-1}\ell_n + p_{n-2}}{q_{n-1}\ell_n + q_{n-1}} \, .$$

6.1. Proposition. *In these circumstances, if $P = p_{n-1}$ and $Q = q_{n-1}$ then $p > p_\circ > 0$, $q > q_\circ > 0$, and*

$$|Pq - pQ| = 1 \, .$$

If (P, Q) is any other solution of this equation then $(P, Q) = (p_{n-1}, q_{n-1}) + k(p, q)$ for some integer k.

Remark. The sequence for a rational number is also repetitive in some sense. One way to interpret this is that if λ is rational, we eventually pass off to $1/0 = \infty$, which is a fixed point of the process.

∘ ———— ∘

And here is the algorithm spelled out in detail:

def euclid(n, m):
  $p_2 = 0$
  $p_1 = 1$
  $q_2 = 1$
  $q_1 = 0$
  while $m \neq 0$:
    $q = n/m$
    $r = n - m \cdot q$
    $n = m$
    $m = r$
    $p = p_1 \cdot q + p_2$
    $q = q_1 \cdot q + q_2$
    $p_2 = p_1$
    $p_1 = p$
    $q_2 = q_1$
    $q_1 = q$
  return $[p_2, q_2]$

7. Characterization of convergents

Suppose

$$\lambda = \langle\!\langle \ell_0, \dots, \ell_{n-1}, \ell_n, \ell_{n+1}, \dots \rangle\!\rangle$$

to be the continued fraction expression of an irrational λ. Then

$$\lambda = \frac{p_{n-1}\lambda_n + p_{n-2}}{q_{n-1}\lambda_n + q_{n-2}}$$

for every n, in which

$$\lambda_n = \langle\!\langle \ell_n, \ell_{n+1}, \dots \rangle\!\rangle \,.$$

It is then true that

$$\frac{p_{n-1}}{q_{n-1}} = \langle\!\langle \ell_0, \dots, \ell_{n-1} \rangle\!\rangle \,,$$

but *this is not necessarily the continued fraction expansion of* p_n/q_n. For one thing, it cannot be this if $\ell_{n-1} = 1$. It is in fact the alternative with parity $(-1)^n$ mentioned in Proposition 3.7.

7.1. Proposition. *Suppose*

$$\lambda = \frac{p\mu + p_\circ}{q\mu + q_\circ} \,,$$

with $\mu > 1$, $q > q_\circ > 0$, *and*

$$pq_\circ - p_\circ q = \pm 1 \,.$$

Then $p_\circ/q_\circ$ *and* p/q *are successive convergents of* λ.

Proof. It is quite explicit.

Let $\varepsilon = pq_\circ - p_\circ q$ and let

$$(\ell_0, \dots, \ell_{n-1})$$

be the unique array specified by Proposition 3.7 for p/q, such that $\varepsilon = (-1)^n$. Now

$$pq_{n-1} - p_{n-1}q = pq_\circ - p_\circ q, \quad p(q_{n-1} - q_\circ) = q(p_{n-1} - p_\circ) \,.$$

Since p and q are relatively prime, q divides $q_{n-1} - q_\circ$. But $0 < q_{n-1} < q$ and $0 < q_\circ < q$, which implies that $|q_\circ - q_{n-1}| < q$. Hence $q_{n-1} = q_\circ$, and then $p_\circ = p_{n-1}$ as well.

Suppose that

$$\mu = \langle\!\langle m_0, m_1, m_2, \dots \rangle\!\rangle \,.$$

then $m_0 \geq 1$ since $\mu > 1$, so

$$\lambda = \langle\!\langle \ell_0, \dots, \ell_{n-1}, \mu_0, \mu_1, \dots \rangle\!\rangle \,.$$

I'll say that two numbers λ and μ are **equivalent** if

$$\lambda = \frac{a\mu + b}{c\mu + d}$$

for some

$$\Pi = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

in $\mathrm{GL}_2(\mathbb{Z})$. I'll call them **properly equivalent** if Π can be chosen to have determinant 1.

7.2. Corollary. *Two irrational numbers λ and μ are equivalent if and only if*

$$\lambda = \langle\!\langle a_0, \ldots, a_{m-1}, c_0, c_1, \ldots \rangle\!\rangle$$
$$\mu = \langle\!\langle b_0, \ldots, b_{n-1}, c_0, c_1, \ldots \rangle\!\rangle$$

for some common tail $(c_0, c_1, \ldots)$.

One can choose the matrix Π *so that* $\det(\Pi) = (-1)^{m-n}$.

Proof. One way is immediate. For the other, suppose that

$$\lambda = \frac{a\mu + b}{c\mu + d} = \Pi(\mu)$$

for some

$$\Pi = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

in $\mathrm{GL}_2(\mathbb{Z})$. Changing signs if necessary, we may assume the denominator to be positive. Suppose that

$$\mu = \langle\!\langle c_0, \ldots, c_{k-1}, c_k, c_{k+1}, \ldots \rangle\!\rangle .$$

The claim is that if k is large enough then there exists $(a_0, a_1, \ldots, a_{m-1})$ such that

$$\lambda = \langle\!\langle a_0, a_1, \ldots, a_{m-1}, c_k, c_{k+1}, \ldots \rangle\!\rangle .$$

with $\det(\Pi) = (-1)^{m-k}$.

For each $k \geq 0$ let

$$\mu_k = \langle\!\langle c_k, c_{k+1}, \ldots \rangle\!\rangle ,$$

so that $\mu = \mu_0$. We have then

$$\mu = R_k(\mu_k) \quad \left(R_k = \begin{bmatrix} p_{k-1} & p_{k-2} \\ q_{k-1} & q_{k-2} \end{bmatrix} \right) ,$$

with $\det(R_k) = (-1)^k$. This also gives

$$\lambda = \Pi_k(\mu_k)$$

with

$$\Pi_k = \Pi R_k = \begin{bmatrix} P_{k-1} & P_{k-2} \\ Q_{k-1} & Q_{k-2} \end{bmatrix} .$$

here

$$Q_{k-1} = cp_{k-1} + dq_{k-1}$$
$$Q_{k-2} = cp_{k-2} + dq_{k-2} ,$$

We'll be through if I show that for $k \gg 0$

(7.3) $$Q_{k-1} > Q_{k-2} > 0 .$$

According to Proposition 5.3,

$$p_{k-1} - \mu q_{k-1} = \delta / q_{k-1}, \quad p_{k-1} - \mu q_{k-2} = \varepsilon / q_{k-2}$$

with $|\delta|$, $|\varepsilon| < 1$. Therefore

$$\begin{aligned} Q_{k-1} &= cp_{k-1} + dq_{k-1} \\ &= c(\mu q_{k-1} + \delta/q_{k-1}) + dq_{k-1} \\ &= (c\mu + d)q_{k-1} + c\delta/q_{k-1} \\ Q_{k-2} &= cp_{k-2} + dq_{k-2} \\ &= c(\mu q_{k-2} + \delta/q_{k-2}) + dq_{k-2} \\ &= (c\mu + d)q_{k-2} + c\delta/q_{k-2} . \end{aligned}$$

But $c\mu + d > 0$, $q_{k-1} > q_{k-2} > 0$, and q_{k-1} and q_{k-2} both grow to ∞ as k does. This implies that (7.3) holds for large k. ∎

I'll show elsewhere how these two results allow us to compute the units and the ideal classes in quadratic extensions of $\mathbb{Q}$.

8. References

1. Harold Davenport, **The higher arithmetic** Cambridge University Press, sixth edition, 1992.

2. Leonhard Euler, 'De fractionibus continuis dissertatio', *Comm. Acad. Sci. Petropol.* **9** (1744), 98–137. Also in **Opera Omnia, ser. I 14**, 187–215. Translation into English by Myra and Bostwick Wyman, 'An essay on continued fractions', *Mathematical Systems Theory* **18** (1985), 295–328.

3. G. H. Hardy and E. M. Wright, **An introduction to the theory of numbers** (fourth edition), Oxford University Press, 1960.

4. Patrick Popescu-Pampu, 'The geometry of continued fractions and the topology of surface singularities', preprint, 2005. Available at

`https://arxiv.org/abs/math/0506432`

This offers a different way of seeing the geometry involved in continued fractions.

5. Henry J. Smith, 'Note on continued fractions', *Messenger of Mathematics* **vi** (1876), 1–14.

An interesting discussion of the relationship between continued fractions and transcendality can be found at

`https://mathoverflow.net/questions/24958/showing-e-is-transcendental-using-its-continued-fraction-expansion`