

MA247 Mathematical Excursions
Approximation of Irrationals by Rationals

T.J. Sullivan
University of Warwick
March 2002

INTRODUCTION

In this essay we shall consider various means by which we can approximate a given real (usually irrational) number by a quotient p/q for suitable integers p and q . Our primary practical motivation, to which we shall refer from time to time, is that computers are limited to a finite subset of the rational numbers: it would be interesting to know if we can find representations for irrationals other than the radix 2 expansions used by computers. We shall also consider the classic problems of the reflected ray and planetary occultations.

We shall consider methods of approximation derived from the theory of continued fractions and the theorems of Dirichlet and Kronecker, which bound the error in our approximations and guarantee more general approximations respectively.

Unless otherwise indicated a proof is a paraphrase of the quoted original.

APPROXIMATION OF IRRATIONALS BY RATIONALS

We shall first introduce some notation:

Definitions 1: The *floor* of $x \in \mathbb{R}$, written $\lfloor x \rfloor$, is the greatest integer less than or equal to x :

$$\lfloor x \rfloor = \sup\{n \in \mathbb{Z} \mid n \leq x\}$$

The *integer part* of x is

$$[x] = \lfloor |x| \rfloor \operatorname{sgn} x,$$

where

$$\operatorname{sgn} x = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}.$$

The *fractional part* of x is defined to be

$$\mathfrak{F}(x) = |x - [x]|.$$

So $x = |x| \operatorname{sgn} x = [x] + \mathfrak{F}(x) \operatorname{sgn}(x)$.

We define x rounded (or the round of x) to be the ‘nearest’ integer to x (rounding $\pm(n+1/2)$ ‘up,’ i.e. away from 0):

$$\mathfrak{R}(x) = \begin{cases} [x] & 0 \leq \mathfrak{F}(x) < 1/2 \\ [x] + \text{sgn } x & 1/2 \leq \mathfrak{F}(x) < 1 \end{cases}$$

Definitions 2: [1]. If we have a set S and a point x we say that x is a *limit point* of S if every neighbourhood of x contains at least one point of S that is not x . We define the *derived set* of S , often written S' , to be the set of all limit points of S . The *closure* of S is $\text{Cl}(S) = S \cup S'$.

As a simple example, $\text{Cl}((0,1)) = (0,1)' = [0,1]$.

Definition 3: A subset S is *dense* in a topological space T if $\text{Cl}(S) = T$, i.e. there is at least one $s \in S$ within every neighbourhood of every point of T .

Theorem 1: \mathbb{Q} is dense in \mathbb{R} .

Proof: (Author’s own proof.) Let $x \in \mathbb{R}$ and $\varepsilon > 0$. It follows immediately from the definition of $[x]$ that

$$0 \leq |x - [x]| < 1,$$

with equality on the left if and only if $x \in \mathbb{Z}$. Let $n > -\log_2 \varepsilon$, so $1/2^n < \varepsilon$. Then

$$\begin{aligned} & 0 \leq |x2^n - [x2^n]| < 1 \\ \Rightarrow & 0 \leq |x2^n - [x2^n]|/2^n < 1/2^n \\ \Rightarrow & 0 \leq |x - [x2^n]/2^n| < \varepsilon \end{aligned}$$

and $[x2^n]/2^n \in \mathbb{Q}$. So \mathbb{Q} is dense in \mathbb{R} . ■

Theorem 1 guarantees that we can find a rational number providing as good an approximation as we like to any real number. What is of much more interest is how accurately we can approximate a given real number subject to certain constraints on the rationals p/q that we may use in our approximation, e.g. $q \leq 1000$, say. The practical interest of this lies mainly in computation: computers are limited to a finite number set, \mathbb{F} , the floating-point numbers.[†] The accuracy with which a computer can approximate a given real is thus clearly limited; it would be useful to know exactly how limited.

[†] See Appendix – Floating-Point Number Systems

As stated in the introduction, we shall consider this problem from two largely disjoint points of view: the theory of continued fractions and the theorems of Dirichlet and Kronecker.

Definition 4: [2: IV, p.4]. In order to save space we shall write

$$q_0 + \frac{1}{q_1 + \frac{1}{q_2 + \dots \frac{1}{q_t}}}$$

for the finite continued fraction

$$q_0 + \frac{1}{q_1 + \frac{1}{q_2 + \frac{1}{\ddots + \frac{1}{q_t}}}}$$

where $q_0 \in \mathbb{Z}$ and $q_i \in \mathbb{N}$ for $i \geq 1$.

Proposition 2: [2: IV, p.6]. *Let $a \in \mathbb{Z}$ and $b \in \mathbb{N}$ and let q_i be as in the Euclidean Algorithm for calculating $\text{hcf}\{a, b\}$. Then*

$$\frac{a}{b} = q_0 + \frac{1}{q_1 + \dots \frac{1}{q_t}}.$$

Proof: The proof is an easy induction on t and is omitted. ■

Corollary 3: *There is a correspondence between \mathbb{Q} and the set of finite continued fractions.*

This result is not at all surprising, of course, since a finite continued fraction is merely a more involved way of writing a simple vulgar fraction. We now wish to investigate the properties of certain types of infinite continued fractions, i.e. those continued fractions for which the sequence (q_i) does not terminate. Before we do so we shall require some new notation.

Theorem 4: [2: IV, pp.9-10]. *Let $\langle q_0, q_1, \dots, q_t \rangle$ denote the polynomial in q_0, q_1, \dots, q_t which is the numerator of the finite continued fraction $q_0 + \frac{1}{q_1 + \dots \frac{1}{q_t}}$. Then*

(i) $\langle q_0, q_1, \dots, q_t \rangle$ is defined by the recurrence relation

$$\langle \rangle = 1,$$

$$\begin{aligned}\langle q_0 \rangle &= q_0, \\ \langle q_0, \dots, q_t \rangle &= q_0 \langle q_1, \dots, q_t \rangle + \langle q_2, \dots, q_t \rangle.\end{aligned}$$

$$(ii) \quad q_0 + \frac{1}{q_1 + \frac{1}{q_2 + \dots + \frac{1}{q_t}}} = \frac{\langle q_0, \dots, q_t \rangle}{\langle q_1, \dots, q_t \rangle}.$$

Proof: We use induction on t . Clearly both statements hold for $t = 0$. Assume them up to the case $t = k$ for some integer $k \geq 1$. Define

$$\begin{aligned}x &= q_0 + \frac{1}{q_1 + \frac{1}{q_2 + \dots + \frac{1}{q_k}}}, \\ y &= q_1 + \frac{1}{q_2 + \frac{1}{q_3 + \dots + \frac{1}{q_k}}}.\end{aligned}$$

Note that $x = q_0 + 1/y$. By our induction hypothesis,

$$y = \frac{\langle q_1, \dots, q_k \rangle}{\langle q_2, \dots, q_k \rangle},$$

and so

$$x = q_0 + \frac{\langle q_2, \dots, q_k \rangle}{\langle q_1, \dots, q_k \rangle} = \frac{q_0 \langle q_1, \dots, q_k \rangle + \langle q_2, \dots, q_k \rangle}{\langle q_1, \dots, q_k \rangle} = \frac{\langle q_0, \dots, q_k \rangle}{\langle q_1, \dots, q_k \rangle},$$

which proves both parts. ■

We now move into the realm of infinite continued fractions, ones for which the sequence of denominators (q_i) does not terminate. We extend our notation for finite continued fractions in the obvious ways, and introduce the notion of convergents:

Definition 5: [2: IV, p.15]. We define

$$\begin{aligned}A_n &= \langle q_0, \dots, q_n \rangle, \\ B_n &= \langle q_1, \dots, q_n \rangle,\end{aligned}$$

and call the ratio A_n/B_n the n^{th} *convergent* of the continued fraction $q_0 + \frac{1}{q_1 + \dots}$.

By Theorem 4 we know that

$$\begin{aligned}A_n &= q_n A_{n-1} + A_{n-2}, \\ B_n &= q_n B_{n-1} + B_{n-2}.\end{aligned}$$

We now present and prove a method for forming a sequence of continued fraction approximations to a given $\alpha \in \mathbb{R}$.

Theorem 5: [2: IV, pp.24-25]. Let $\alpha \in \mathbb{R}$. We define the integers q_i by

$$\alpha_0 = \alpha, q_0 = [\alpha_0],$$

$$\alpha_i = 1/(\alpha_{i-1} - q_{i-1}), q_i = [\alpha_i] \text{ for } i \geq 1.$$

Then the finite continued fraction $q_0 + \frac{1}{q_1 + \frac{1}{q_2 + \frac{1}{\dots + \frac{1}{q_n}}}}$ converges to α as $n \rightarrow \infty$.

Proof: (Author's own proof, abridged.) Since $q_i \in \mathbb{N}$ for $i \geq 1$, the recursion formula

$$B_n = q_n B_{n-1} + B_{n-2}$$

tells us that the sequence $(B_i)_{i=1}^\infty$ is a strictly increasing sequence of integers, and hence tends to ∞ as $n \rightarrow \infty$. We can easily show that

$$\alpha = \frac{\langle q_0, \dots, q_n, \alpha_{n+1} \rangle}{\langle q_1, \dots, q_{n+1}, \alpha_{n+2} \rangle} = \frac{\alpha_{n+1} A_n + A_{n-1}}{\alpha_{n+1} B_n + B_{n-1}}$$

and hence (with a little algebraic manipulation) that

$$\left| \alpha - \frac{A_n}{B_n} \right| < \frac{1}{B_n B_{n+1}} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

■

This method of approximation of irrationals can be quite useful, and confirms many of our previous guesses about certain irrationals. For example, the rational approximation $22/7$ (known to the ancient Greeks) is often used for π . Using the method above, we see that $22/7$ is the first convergent to π . (The second, third and fourth convergents are $333/106$, $355/113$ (known to the ancient Chinese) and $103993/33102$ respectively.)

We can easily perform many algebraic operations on continued fractions. So, we might choose to record π not as a radix expansion $\pi = 3.141\dots$ (up to the accuracy of our computer's floating-point system) but as a sequence of integers i.e. the q_i in the method above. This is not a method that is used very frequently; it is included here only for academic interest.

Definitions 6: [2: IV, p.31]. We say that $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ is a *quadratic irrational* if it is an irrational root of a quadratic equation with integer coefficients; that is, $\alpha = x + y\sqrt{D}$ for $x, y \in \mathbb{Q}$ and $D \in \mathbb{N}$ not a perfect square. We call $\alpha' = x - y\sqrt{D}$ the *conjugate* of

α . Furthermore, if $\alpha > 1$ and $-1 < \alpha' < 0$ then we say that α is a *reduced quadratic irrational*.

Definitions 7: We say that a continued fraction $q_0 + \frac{1}{q_1 + \dots}$ is *periodic* if there exist $N \in \mathbb{N} \cup \{0\}$ and $L \in \mathbb{N}$ such that $n \geq N \Rightarrow q_n = q_{n+L}$, i.e. the q_i eventually repeat. If the least such $N = 0$ we say that the continued fraction is *purely periodic* or *periodic without delay*.

We now state without proof two theorems concerning the continued fraction representations of quadratic irrationals. The proofs are too long to give here.

Theorem 6: [2: IV, p.38]. *A real number has a purely periodic representation as a continued fraction if and only if it is a reduced quadratic irrational.*

Theorem 7: [2: V, p.5]. *A real number has a periodic representation as a continued fraction if and only if it is a quadratic irrational.*

The general idea of the proofs is that if one were to follow the method of Theorem 5 one would find that the discriminant D of the quadratic irrational $\alpha = x + y\sqrt{D}$ would also be the discriminant of later α_i , hence the periodicity. The converse arguments are not at all difficult.

The implication of Theorem 7 is quite strong in answering our question about how we may represent irrational numbers. Note that if a given $\alpha \in \mathbb{R}$ is a quadratic irrational then it can be calculated to an arbitrary degree of accuracy knowing only a finite number of integers – the first N of the ‘delay’ and the L of the period. Provided these integers lie in our floating-point set \mathbb{F} we can now compute any quadratic irrational to any degree of accuracy we desire. Again, this method is more of academic than practical interest.

We now turn our attention to a theorem of Dirichlet’s that bounds the error involved in our approximation of an arbitrary real by a suitable rational.

Theorem 8: (Dirichlet’s Theorem) [3: p.155 and §11.3, pp.156-157]. *For any $\theta \in \mathbb{R}$ there is at least one rational p/q satisfying*

$$\left| \frac{p}{q} - \theta \right| < \frac{1}{q^2} \quad (*)$$

Moreover, if $\theta \in \mathbb{R} \setminus \mathbb{Q}$ then there is an infinity of solutions to ().*

Proof: If $\theta \in \mathbb{Q}$ then the proof is trivial, as $\theta = p/q$ for some $p, q \in \mathbb{Z}$ by definition of rationality. So, assume $\theta \in \mathbb{R} \setminus \mathbb{Q}$ and fix $\varepsilon > 0$ and $Q \in \mathbb{N}$. The $Q+1$ points in the set $\{\mathfrak{F}(r\theta) \mid r = 0, 1, \dots, Q\}$ all lie in the interval $[0, 1)$. It is clear that at least two

must lie in an interval of the form $[s/Q, (s+1)/Q)$, since there are Q of these and they partition $[0,1)$. So there exist integers $q_1, q_2 \leq Q$ such that $|\mathfrak{F}(q_1\theta) - \mathfrak{F}(q_2\theta)| < 1/Q$.

We may assume without loss of generality that $q_1 < q_2$; set $q = q_2 - q_1$. Then $0 < q \leq Q$ and $|q\theta - \mathfrak{R}(q\theta)| < 1/Q$. Hence, there is an integer p such that $|q\theta - p| < 1/Q$, and so

$$\left| \frac{p}{q} - \theta \right| < \frac{1}{qQ}. \quad (**)$$

Now let $Q = \lfloor 1/\varepsilon \rfloor + 1$. By the argument above, for all $\varepsilon > 0$ there exists a $q \leq \lfloor 1/\varepsilon \rfloor + 1$ such that

$$\left| \frac{p}{q} - \theta \right| < \frac{\varepsilon}{q} \leq \frac{1}{q^2}.$$

From here it is simple to show that the number of solutions to inequality (*) is infinite when θ is irrational: suppose that $\{p_i/q_i \mid i = 1, \dots, t\}$ contains all the solutions. Since θ is irrational there is a Q such that, for $i = 1, \dots, t$,

$$\left| \frac{p_i}{q_i} - \theta \right| > \frac{1}{Q}.$$

But then the p/q of (**) satisfies

$$\left| \frac{p}{q} - \theta \right| < \frac{1}{qQ} \leq \frac{1}{Q}$$

and is not one of the p_i/q_i , which is a contradiction. Hence, there is an infinity of solutions to (*) when θ is irrational. ■

Note that the proof of Theorem 5 showed that the convergents of a continued fraction satisfy inequality (*).

Dirichlet's Theorem may be generalized to situations in which we attempt to simultaneously approximate k reals by appropriate rationals.

Theorem 9: (Dirichlet's Theorem in k dimensions.) [3: p.170]. *Let $\theta_1, \dots, \theta_k \in \mathbb{R}$. Then there exist $p_1, \dots, p_k, q \in \mathbb{Z}$ such that, for each $i = 1, \dots, k$,*

$$\left| \frac{p_i}{q} - \theta_i \right| < \frac{1}{q^{1+\mu}},$$

where $\mu = 1/k$. If any of the θ_i are irrational then there is an infinity of solutions to the above.

Note that in some sense the accuracy of our approximations decreases as we approximate more numbers.

Proof: We can assume without loss of generality that all $\theta_i \in [0,1)$. Pick $Q \in \mathbb{N}$ and consider the ‘cube’ $[0,1)^k$ partitioned into Q^k ‘boxes’ by planes drawn parallel to each face, at distances $1/Q$. Of the $Q^k + 1$ points in

$$\{(\mathfrak{F}(r\theta_1), \dots, \mathfrak{F}(r\theta_k)) \mid r = 0, 1, \dots, Q^k\} \subset [0,1)^k$$

some two must lie in the same box – those corresponding to $r = q_1$ and $r = q_2 > q_1$, say. As in the proof of Theorem 8 set $q = q_2 - q_1$. Then there always exists a $q \leq Q$ such that, for all $i = 1, \dots, k$,

$$|q\theta_i - \mathfrak{R}(q\theta_i)| < \frac{1}{Q} \leq \frac{1}{q^\mu}.$$

If a particular θ_i is irrational then write θ_i for θ in the proof of Theorem 8. ■

Corollary 10: [3: p.170]. Given $\theta_1, \dots, \theta_k$ and $\varepsilon > 0$ we can find $q \in \mathbb{Z}$ such that $q\theta_i$ differs from an integer by less than ε for each $i = 1, \dots, k$.

We now turn our attention to Kronecker’s Theorem, which is similar to Dirichlet’s Theorem, but is somewhat stronger because of the inclusion of an arbitrary $\alpha \in \mathbb{R}$. We previously considered the question, ‘Given θ can we find an integer n such that $n\theta$ is nearly an integer?’ We now ask, ‘Given θ and α can we find an integer n such that $n\theta - \alpha$ is nearly an integer?’

Theorem 11: (Kronecker’s Theorem in 1 dimension) [3: pp.375-376]. If θ is irrational, $\alpha \in \mathbb{R}$ and $N, \varepsilon > 0$ then there exist integers $n > N$ and p such that

$$|n\theta - p - \alpha| < \varepsilon.$$

Equivalently:

If θ is irrational then the set $\{\mathfrak{F}(r\theta) \mid r \in \mathbb{N}\}$ is dense in the unit interval $[0,1]$.

Proof: Corollary 10 with $k = 1$ tells us that there exist integers n_1 and p such that

$$|n_1\theta - p| < \varepsilon. \tag{*}$$

Hence, it is true that either $|\mathfrak{F}(n_1\theta)| < \varepsilon$ or $|\mathfrak{F}(n_1\theta) - 1| < \varepsilon$.

Define the sequence $(a_r)_{r=0}^\infty$ by $a_r = \mathfrak{F}(rn_1\theta)$. The points of this sequence lie entirely in $[0, 1]$. By (*), the fact that $\mathfrak{F}(kx) = \mathfrak{F}(k\mathfrak{F}(x))$ and the assumption that $\theta \in \mathbb{R} \setminus \mathbb{Q}$, we see that for all $\alpha \in [0, 1]$ there is an a_r such that $|\alpha - a_r| < \varepsilon$. ■

This is just one of many proofs of Kronecker's Theorem in 1 dimension. It can also be shown (though we shall not do so here) that the points $\mathfrak{F}(r\theta)$ are uniformly distributed in $(0, 1)$.

We shall now apply Kronecker's Theorem in 1 dimension to a geometrical problem of some interest in the field of dynamical systems. The problem was first solved by König and Szücs in 1913.

The Reflected Ray Problem [3: pp.378-380]

Consider the following setup: a ray of light starts at a position $P = (a, b) \in (-1/2, 1/2)^2 = S$ with velocity $(u, v) \in \mathbb{R}^2$. The boundary of the square S is a perfect mirror (we assume that if the ray strikes a corner it returns back along the path it took previously, as is suggested by considerations of continuity). What is the nature of the ray's path?

Proposition 12: *The path of the ray is either closed and periodic or dense in the square S . The path is closed and periodic if and only if u/v is rational.*

Proof: [4] in [3: pp.378-380]. Let Λ denote the path of the ray inside the square S . Consider all reflections of P in the sides of S , as shown in the diagram below:

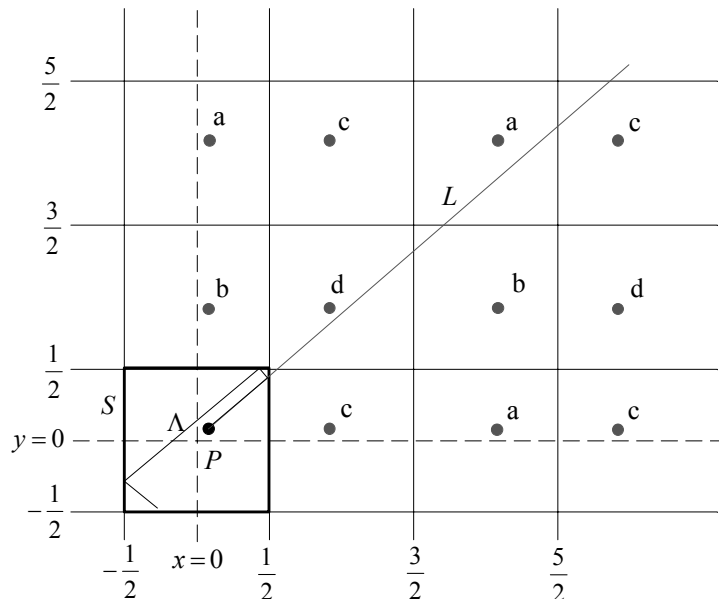


Fig. 1 – The reflected ray, Λ , and the continued ray, L .

Note that there are four types of image of P , those with co-ordinates

- a. $(a + 2m, b + 2n)$,
- b. $(a + 2m, -b + 2n + 1)$,
- c. $(-a + 2m + 1, b + 2n)$,
- d. $(-a + 2m + 1, -b + 2n + 1)$,

for some $m, n \in \mathbb{Z}$. In each case the image of the velocity is

- a. (u, v) ,
- b. $(u, -v)$,
- c. $(-u, v)$,
- d. $(-u, -v)$.

Consider now the path the ray would take if the mirrors were absent: this would be a straight half-line, L , say. If we tile the plane with reflections of the square S we see that each segment of L in an image of S corresponds to exactly one portion of the path Λ . Moreover, Λ will be closed and periodic if and only if L passes through a type-a image of P . Since $L = \{(a + ut, b + vt) \mid t \in \mathbb{R}, t \geq 0\}$, Λ will be closed and periodic if and only if there are integers m, n such that $(ut, vt) = (2m, 2n)$, i.e. u/v is rational.

We now need to show that if u/v is irrational then Λ is dense in S . Let $(\xi, \eta) \in S$ be arbitrary. Λ is dense in S if and only if L passes arbitrarily close to some image of (ξ, η) , which occurs if L passes arbitrarily close to some type-a image of (ξ, η) , which occurs if there exist $t > 0$ and $m, n \in \mathbb{Z}$ such that, for all $\varepsilon > 0$,

$$\begin{aligned} |a + ut - \xi - 2m| &< \varepsilon, \\ |b + vt - \eta - 2n| &< \varepsilon. \end{aligned}$$

Let $t = (\eta + 2n - b)/v$ so that the second inequality is satisfied. Now let $\theta = u/v$ and $\omega = \theta \frac{b - \eta}{2} - \frac{a - \xi}{2}$. The first inequality then becomes

$$|n\theta - m - \omega| < \varepsilon/2,$$

which Theorem 11 assures us is true. So Λ is dense in S . ■

Note the strength of this statement: we are not merely asserting that the projections of Λ on to the x - and y -axes are each dense in $(-1/2, 1/2)$, we are saying that Λ itself

is dense in the *entire square* S . This distinction is important. For example, if $u = v$ then the path Λ will be closed and periodic, crossing itself if and only if $a = b$, and yet its projections onto the two axes will each be dense in the appropriate intervals.

This problem is also, essentially, a single induction step in the proof of Kronecker's Theorem in its general form, Theorem 13. After reading Definition 8 the reader may see that the statement ' u/v is irrational' is in fact equivalent to the statement ' $1, u, v$ are linearly independent over \mathbb{Z} ,' which will make the connection even clearer.

We shall see a more general (but slightly different) version of this problem later; the reader may wish to consider the nature of the path of a light ray in a cube, or a higher-dimensional hypercube.

Definition 8: We say that x_1, \dots, x_n are *linearly independent over \mathbb{Z}* if no relation of the form

$$\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_n x_n = 0$$

with $\lambda_i \in \mathbb{Z}$ holds unless all $\lambda_i = 0$.

Theorem 13: (Kronecker's Theorem in k dimensions) [3: p.382]. *If $1, \theta_1, \dots, \theta_k$ are linearly independent over \mathbb{Z} , $\alpha_1, \dots, \alpha_k$ are arbitrary and $N, \varepsilon > 0$ then there exist integers $n > N$ and p_1, \dots, p_n such that, for $i = 1, \dots, k$,*

$$|n\theta_i - p_i - \alpha_i| < \varepsilon.$$

Equivalently:

If $1, \theta_1, \dots, \theta_k$ are linearly independent over \mathbb{Z} then the set

$$\{(\mathfrak{F}(r\theta_1), \dots, \mathfrak{F}(r\theta_k)) \mid r \in \mathbb{N}\}$$

is dense in the unit cube $[0, 1]^k$.

Proof: [5] in [3: pp.386-388]. This proof was first given by Estermann – there are also proofs of quite different character by Lettenmeyer (who takes a geometrical approach) and Bohr (who uses analytical methods). Estermann's proof is of the first form of the theorem and uses induction on k . Very roughly, the idea is to make the error in the approximation of the first $k - 1$ numbers small, and show that the error introduced by the k^{th} is also small, so that the total error is also small.

In fact, Estermann's proof is of the slightly stronger statement that if $1, \theta_1, \dots, \theta_k$ are linearly independent over \mathbb{Z} , $\alpha_1, \dots, \alpha_k$ are arbitrary, $\lambda \in \mathbb{R} \setminus \{0\}$ and $\omega, \varepsilon > 0$ then there exist integers n and p_1, \dots, p_n such that, for $i = 1, \dots, k$,

$$\begin{aligned} |n| > \omega, \operatorname{sgn} n = \operatorname{sgn} \lambda, \\ |n\theta_i - p_i - \alpha_i| < \varepsilon.^\ddagger \end{aligned}$$

By Corollary 10 there are integers $s > 0$ and b_1, \dots, b_k such that, for $i = 1, \dots, k$,

$$|s\theta_i - b_i| < \varepsilon/2.$$

Since the θ_i are irrational $s\theta_k - b_k \neq 0$. If we define for $i = 1, \dots, k$

$$\phi_i = \frac{s\theta_i - b_i}{s\theta_k - b_k}$$

then the linear independence over \mathbb{Z} of the θ_i and 1 implies that the ϕ_i are also linearly independent over \mathbb{Z} . Note that $\phi_k = 1$.

Suppose now that $k \geq 2$ and assume the theorem is true for the case $k - 1$. Consider

$$\begin{aligned} \phi_1, \dots, \phi_{k-1} \text{ for the } \theta_i, \\ \beta_1, \dots, \beta_{k-1}, \text{ where } \beta_i = \alpha_i - \alpha_k \phi_i, \text{ for the } \alpha_i, \\ \varepsilon/2 \text{ for } \varepsilon, \\ \lambda(s\theta_k - b_k) \text{ for } \lambda, \\ \Omega = (\omega + 1)|s\theta_k - b_k| + |\alpha_k| \text{ for } \omega. \end{aligned}$$

The theorem tells us that there are integers c_k, c_1, \dots, c_{k-1} such that, for $i = 1, \dots, k - 1$,

$$\begin{aligned} |c_k| > \Omega, \operatorname{sgn} c_k = \operatorname{sgn} \lambda(s\theta_k - b_k), \\ |c_k \phi_i - c_i - \beta_i| < \varepsilon/2. \end{aligned}$$

The last inequality may be rewritten to hold for all $i = 1, \dots, k$ as

$$\left| \frac{c_k + \alpha_k}{s\theta_k - b_k} (s\theta_i - b_i) - c_i - \alpha_i \right| < \varepsilon/2.$$

This inequality is clearly true for $k = 1$ once we choose a c_k suitably large in absolute value.

Pick an integer N such that

$$\left| N - \frac{c_k + \alpha_k}{s\theta_k - b_k} \right| < 1$$

[‡] Estermann's version reduces to the original if λ is positive.

and let $n = Ns$ and $p_i = Nb_i + c_i$. Then, for each $i = 1, \dots, k$,

$$\begin{aligned} |n\theta_i - p_i - \alpha_i| &= |N(s\theta_i - b_i) - c_i - \alpha_i| \\ &\leq \left| \frac{c_k + \alpha_k}{s\theta_k - b_k} (s\theta_i - b_i) - c_i - \alpha_i \right| + |s\theta_i - b_i| \\ &< \varepsilon/2 + \varepsilon/2 \\ &= \varepsilon \end{aligned}$$

Now,

$$\left| \frac{c_k + \alpha_k}{s\theta_k - b_k} \right| \geq \frac{|c_k| - |\alpha_k|}{|s\theta_k - b_k|} > \omega + 1,$$

so $|N| > \omega$ and so $|n| = |Ns| \geq |N| > \omega$. Also, $\text{sgn } n = \text{sgn } N = \text{sgn } \frac{c_k}{s\theta_k - b_k} = \text{sgn } \lambda$.

This completes the induction from $k-1$ to k . Hence, the theorem is true for all $k \in \mathbb{N}$. ■

Planetary Occultations [3: §23.6, p.384]

Kronecker's Theorem in k dimensions may be used to examine the following problem. Suppose we have a system of k planets, P_1, \dots, P_k , orbiting a star at the origin, O , these orbits being circular and coplanar. Suppose that the diameters of the planets are such that when viewed from O each planet P_i will completely obscure (occult) P_j if (O, P_i, P_j) are collinear and $i < j$.

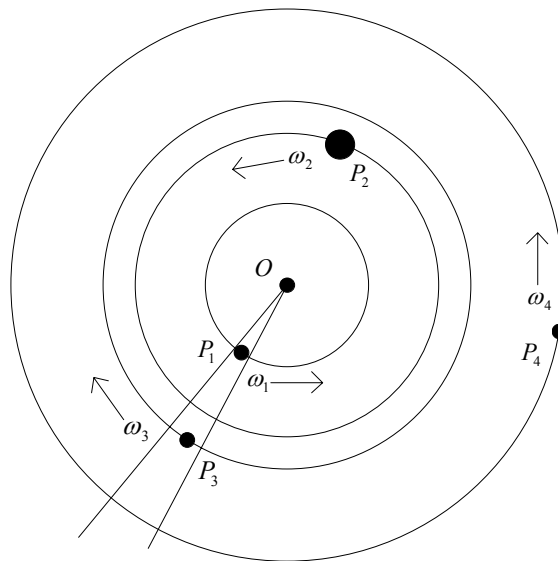


Fig. 2 – A four-planet system. P_1 occults P_3 .

Given a reference line, the planets' initial angular displacements from this line, α_i , and their angular velocities, ω_i , can we determine whether total planetary occultations (P_1 occults all the other P_i , i.e. all the planets' positions are as close to being collinear as we please) will never occur, occur only finitely many times, or infinitely many times?

Let us first consider the case in which $\alpha_i = 0$ for all $i = 1, \dots, k$. At an arbitrary time t planet P_i has angular displacement $t\omega_i$ (modulo 2π). Corollary 10 tells us that we can find infinitely many t such that, for each $i = 1, \dots, k$, $\mathfrak{F}(t\omega_i/2\pi)$ is as close as we wish to 0 (or 1, but $0 \equiv 1$ (modulo 1)). So, regardless of the ω_i , we shall get infinitely many total occultations.

If the α_i are not all 0 then occultations may not occur at all; this is where the requirement on linear independence over \mathbb{Z} comes into play. For instance, if $\omega_1 = \omega_2$ and (O, P_1, P_2) are not collinear at $t = 0$ then they will never be collinear. However, Theorem 13 tells us that if the ω_i and 1 are linearly independent over \mathbb{Z} then we still get infinitely many total planetary occultations.

CONCLUSION

We have shown that we can, in theory, approximate any real number arbitrarily well by a suitable rational number (Theorem 1), and have given a method for forming an increasingly accurate sequence of approximations (Theorem 5).

One of our practical motivations was to study ways in which we might store or compute irrational numbers using a finite number set \mathbb{F} , the floating-point numbers. Although the standard method of using a base 2 radix expansion is quite accurate, we have seen that, in certain circumstances, we can use finite sequences of integers to represent irrationals to arbitrary degrees of accuracy (Theorems 6 and 7).

We have seen how Dirichlet's Theorem determines bounds on the error in our approximations, both for approximation of a single number (Theorem 8) and for several numbers simultaneously (Theorem 9). Finally, we have shown how Kronecker's Theorem (Theorem 13) guarantees more general approximation and may be applied to solve the problems of the reflected ray and planetary occultations.

Some of the theory developed above, in particular Kronecker's Theorem, is of considerable relevance to the study of dynamical systems. For example, in [6: §2.3, p.40], Moser is making use of Kronecker's Theorem in its general case when he says that if a flow on an k -dimensional torus is given by the frequencies $\omega_1, \dots, \omega_k$ then the flow is dense in the torus if the ω_i are "rationally independent" i.e. linearly independent over \mathbb{Z} .

APPENDIX – FLOATING-POINT NUMBER SYSTEMS

A floating-point number is a number like 6.67×10^{-33} :

$$\underbrace{6.67}_{\text{mantissa}} \times \underbrace{10}_{\text{base}}^{\underbrace{-33}_{\text{exponent}}}$$

We have a base b ($=10$), a mantissa m ($=6.67$), $0 \leq m < b$, with a maximum number of digits after the ‘decimal’ point, and an integer exponent e ($=-33$), the choice of which is also bounded above and below. The set of all such numbers is a floating-point set, often denoted by \mathbb{F} .

The first thing that should be clear is that a set of such numbers is finite. For example, if we take $b = 10$, allow up to 2 decimal places and let $-99 \leq e \leq 99$ then we have $9 \times 10 \times 10 \times (2 \times 99 + 1) = 179100$ possible numbers. More generally, if, for some base b , we allow n ‘decimal’ places and $l \leq e \leq u$ then our \mathbb{F} has $b^n(b-1)(l-u+1)$ elements.

Computers make use of floating-point number sets. There is no single standard \mathbb{F} – there are various ones allowing various degrees of accuracy. For example, a ‘long double’ floating-point number used in computing uses 10 bytes (i.e. 80 binary bits) to store a number in absolute value between 3.4×10^{-4932} and 1.1×10^{4932} , together with its sign. It is customary to denote $\sup \mathbb{F}$ as ‘infinity’ and $\inf \mathbb{F}$ as ‘-infinity’.

The IEEE Standard 754: [7]

IEEE (Institute of Electrical and Electronic Engineers) Standard 754 floating-point is the most common representation for real numbers on computers today. IEEE floating-point numbers have three basic components: the sign, the mantissa and the exponent. The base (2) is implicit and is not stored.

The following table shows the layout for single (32-bit) and double (64-bit) precision floating-point numbers. The number of bits for each field are shown (bit ranges are in square brackets):

	Sign	Exponent	Mantissa	Bias
Single Precision	1 [31]	8 [30-23]	23 [22-00]	127
Double Precision	1 [63]	11 [62-52]	52 [51-00]	1023

Table 1 – Single and double precision IEEE floating-point numbers.

Sign: The sign bit denotes the sign of the number: 0 for positive, 1 for negative.

Exponent: Since the exponent field needs to represent both positive and negative exponents a *bias* is added to the actual exponent in order to get the stored exponent.

So, for example, a stored value of 190 corresponds to an actual value of 63. Exponents of all zeros and all ones are reserved for special numbers.

Mantissa: The mantissa stores the precision bits in what is known as *normalized form*. This basically puts the radix point after the first non-zero digit, so 650 would be 6.5×10^2 . (A nice little optimization is available to us with a base of 2: since the only possible non-zero digit is 1 we can toss away the 1 and just assume that it exists, giving us one extra bit of precision for free. Thus, the mantissa has effectively 24 bits of resolution in single precision, and 53 in double precision.)

There are certain special values in IEEE-754:

Zero: 0 is not directly representable in the format described above due to the assumed leading 1 in the mantissa. Zero is a special value denoted with an exponent field of 0s and a mantissa of 0s. There are distinct representations for +0 and -0, but they are treated as being the same.

Denormalized: If the exponent is all 0s, but the mantissa is not (otherwise it would be interpreted as zero) then the value is a *denormalized* number, which does *not* have an assumed leading one before the binary point. One can view zero as a special kind of denormalized number.

Infinity: The values +infinity and -infinity are denoted by an exponent of all 1s and a mantissa of all 0s. The sign bit distinguishes between -infinity and infinity. It is useful to be able to denote infinity as a specific value because it allows operations to continue past what are known as 'overflow situations' in which the numbers are too big or small. Operations with infinite values are well defined in IEEE-754 (see below).

Indeterminate: The value indeterminate is represented by an exponent of all 1s, a mantissa with a leading 1 followed by all 0s, and a sign bit of 1. This value is used to represent results that are indeterminate, such as infinity - infinity, or $0 \times \text{infinity}$.

There is also a value *NaN* (Not a Number) which corresponds to an error of some kind.

Special operations include:

$$\begin{aligned} x/\pm \text{infinity} &= 0, \\ \pm \text{infinity} \times \pm \text{infinity} &= \pm \text{infinity}, \\ \pm x/0 &= \pm \text{infinity}, \\ \text{infinity} + \text{infinity} &= \text{infinity}, \\ \text{infinity} - \text{infinity} &= \text{indeterminate}, \\ \pm \text{infinity}/\pm \text{infinity} &= \text{indeterminate}, \\ \pm \text{infinity} \times 0 &= \text{indeterminate}. \end{aligned}$$

Obviously, any expression involving a value of indeterminate is evaluated as being indeterminate.

BIBLIOGRAPHY

- [1] *Penguin Dictionary of Mathematics*, 2nd Edition, ed. D. Nelson, Penguin (1998)
- [2] *MA246 Number Theory (IV, V)*, T.O. Hawkes, University of Warwick (2001).
- [3] *An Introduction to the Theory of Numbers*, Fifth Edition, G.H. Hardy & E.M. Wright, O.U.P.
- [4] König and Szücs in *Rendiconti del circolo matematico di Palermo*, 36 (1913), 79-90.
- [5] T. Estermann in *The Journal of the London Mathematical Society*, 8 (1933), 18-20.
- [6] *Stable and Random Motions in Dynamical Systems*, Jürgen Moser, Princeton University Press / University of Tokyo Press (1973).
- [7] *IEEE Standard for Binary Floating-Point Arithmetic, IEEE Std. 754-1985*, Institute of Electrical and Electronic Engineers Computer Society (1985).