DIOPHANTINE APPROXIMATIONS

Ivan Niven

DIOPHANTINE APPROXIMATIONS

IVAN NIVEN
Professor of Mathematics
The University of Oregon

THE NINTH ANNUAL SERIES
OF EARLE RAYMOND HEDRICK LECTURES
OF THE MATHEMATICAL ASSOCIATION OF AMERICA

INTERSCIENCE PUBLISHERS
a division of John Wiley & Sons, New York · London

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 63-11444

PRINTED IN THE UNITED STATES OF AMERICA

INTERSCIENCE TRACTS

IN PURE AND APPLIED MATHEMATICS

1. D. Montgomery and L. Zippin—**Topological Transformation Groups**
2. Fritz John—**Plane Waves and Spherical Means Applied to Partial Differential Equations**
3. E. Artin—**Geometric Algebra**
4. R. D. Richtmyer—**Difference Methods for Initial-Value Problems**
5. Serge Lang—**Introduction to Algebraic Geometry**
6. Herbert Busemann—**Convex Surfaces**
7. Serge Lang—**Abelian Varieties**
8. S. M. Ulam—**A Collection of Mathematical Problems**
9. I. M. Gel'fand—**Lectures on Linear Algebra**
10. Nathan Jacobson—**Lie Algebras**
11. Serge Lang—**Diophantine Geometry**
12. Walter Rudin—**Fourier Analysis on Groups**
13. Masayoshi Nagata—**Local Rings**
14. Ivan Niven—**Diophantine Approximations**

Additional volumes in preparation

INTERSCIENCE TRACTS
IN PURE AND APPLIED MATHEMATICS

Number 14

DIOPHANTINE APPROXIMATIONS

By Ivan Niven

INTERSCIENCE PUBLISHERS
a division of John Wiley & Sons, New York · London

PREFACE

At the 1960 summer meeting of the Mathematical Association of America it was my privilege to deliver the Earle Raymond Hedrick lectures. This monograph is an extension of those lectures, many details having been added that were omitted or mentioned only briefly in the lectures. The monograph is self-contained. It does not offer a complete survey of the field. In fact the title should perhaps contain some circumscribing words to suggest the restricted nature of the contents, but such modifiers have been omitted for the sake of simplicity.

The topics covered are: basic results on homogeneous approximation of real numbers in Chapter 1; the analogue for complex numbers in Chapter 4; basic results on non-homogeneous approximation in the real case in Chapter 2; the analogue for complex numbers in Chapter 5; fundamental properties of the multiples of an irrational number, for both the fractional and integral parts, in Chapter 3. Many proofs are offered here for the first time, although the results themselves are not novel.

An attempt has been made, in a section entitled "Further results" at the end of each chapter, to provide a bibliographic account of closely related work. These sections also give the sources from which the proofs are drawn. Having used the literature freely, I wish to acknowledge especially the usefulness of the monographs by Cassels (1956) and Koksma (1936); see the bibliography for detailed references. The monographs by Cassels (1956) and Mahler (1961) treat many topics not considered here, such as the elegant work of Roth on the approximation of algebraic numbers. The topic of Diophantine approximations is also treated to some extent in the general texts on number theory, notably in Hardy and Wright (1960; Chapters 3, 10, 11, 23, 24) and LeVeque (1956; vol. 1, Chapter 9; vol. 2, Chapter 4).

A unique feature of this monograph is that continued fractions are not used. This is a gain in that no space need be given over to their description, but a loss in that certain refinements appear out of reach without the continued fraction approach. Another feature of this monograph is the inclusion of basic results in the complex case, which are often neglected in favor of the real number discussion. The parallel arguments for the real and complex cases in Chapters 2 and 5 are given here for the first time. This development of the theory was rounded out by the Theorem 5.3 of Eggan and Maier, who kindly provided me with their work prior to its publication.

Professor Herbert S. Zuckerman read the entire manuscript and made many helpful suggestions, for which I express my appreciation. I also wish to thank Dr. Charles L. Vanden Eynden for suggesting certain clarifications.

IVAN NIVEN
EUGENE, OREGON

January 1963

NOTATION AND CONVENTIONS

For any real number x:

$[x]$ denotes the greatest integer $\leqq x$; that is, $[x]$ is the unique integer satisfying $[x] \leqq x < [x] + 1$.

$\|x\|$ denotes the absolute value of the difference between x and the nearest integer; thus $\|x\| = \min |x-n|$, where the minimum is taken over all integers n.

(x) denotes the fractional part of x, namely $(x) = x - [x]$; this notation is used only in a few places where it is awkward to write $x - [x]$.

The symbol Z is used to denote the set of all integers.

$a|b$ means that the integer a is a divisor of the integer b.

$\{u_i\}$ denotes the sequence $u_1, u_2, u_3, \cdots$.

$\theta \equiv \lambda \pmod 1$ means that $\theta - \lambda$ is an integer.

N_α denotes the set of the integer parts of the multiples of α, thus

$$[\alpha],\ [2\alpha],\ [3\alpha],\ \cdots.$$

The triangle inequality, $|u + v| \leqq |u| + |v|$, is often used without reference or allusion.

A set $\theta_1, \theta_2, \cdots, \theta_n$ of real numbers is said to be linearly dependent over the rational numbers if there exist rational numbers $r_1, r_2, \cdots, r_n$ not all zero such that $\sum r_j\theta_j = 0$. Note that this is equivalent to saying that there are integers $k_1, k_2, \cdots, k_n$ not all zero such that $\sum k_j\theta_j = 0$. A finite set of real numbers is said to be linearly independent over the rational numbers in case they are not linearly dependent.

Wherever in the text a name appears with a date, such as Koksma (1935), the reference is to the paper or book of that date as listed in the bibliography at the end.

CONTENTS

CHAPTER 1

The Approximation of Irrationals by Rationals

1.1. *The pigeon-hole principle*

Given a real number θ, how closely can it be approximated by rational numbers? To make the question more precise, for any given positive ε is there a rational number a/b within ε of θ, so that the inequality

$$\left|\theta - \frac{a}{b}\right| < \varepsilon$$

is satisfied? The answer is yes because the rational numbers are dense on the real line. In fact, this establishes that for any real number θ and any positive ε there are infinitely many rational numbers a/b satisfying the above inequality.

Another way of approaching this problem is to consider all rational numbers with a fixed denominator b, where b is a positive integer. The real number θ can be located between two such rational numbers, say

$$\frac{c}{b} \leqq \theta < \frac{c+1}{b},$$

and so we have $|\theta - c/b| < 1/b$. In fact, we can write

$$\left|\theta - \frac{a}{b}\right| \leqq \frac{1}{2b} \tag{1}$$

by choosing $a = c$ or $a = c + 1$, whichever is appropriate. The inequality (1) would be strict, that is to say, equality would be excluded if θ were not only real but irrational. We shall confine our attention to irrational numbers θ because most of the questions about approximating rationals by rationals reduce to simple problems in linear Diophantine equations.

Now by use of the pigeon-hole principle (sometimes called the box principle) we can improve inequality (1) as in the following theorem. The pigeon-hole principle states that if $n + 1$ pigeons are in n holes, at least one hole will contain at least two pigeons.

THEOREM 1.1. *Given any irrational number θ and any positive integer m, there is a positive integer $b \leqq m$ such that*

$$\|b\theta\| = |b\theta - a| < \frac{1}{m + 1}.$$

The symbol a here denotes the integer nearest to $b\theta$, so that the equality $\|b\theta\| = |b\theta - a|$ holds by the definition of the symbolism.

Proof: Consider the $m + 2$ real numbers

$$(2) \qquad 0,\ 1,\ \theta - [\theta],\ 2\theta - [2\theta],\ \cdots,\ m\theta - [m\theta]$$

lying in the closed unit interval. Divide the unit interval into $m + 1$ subintervals of equal length

$$(3) \qquad \frac{j}{m + 1} \leqq x < \frac{j + 1}{m + 1}, \qquad j = 0, 1, 2, \cdots, m.$$

Since θ is irrational, each of the numbers (2) except 0 and 1 lies in the interior of exactly one of the intervals (3). Hence two of the numbers (2) lie in one of the intervals (3); thus there are integers k_1, k_2, h_1, and h_2 such that

$$|(k_2\theta - h_2) - (k_1\theta - h_1)| < \frac{1}{m + 1}.$$

We may presume that $m \geqq k_2 > k_1 \geqq 0$. Defining $b = k_2 - k_1$, $a = h_2 - h_1$, we have established the theorem.

Since $(m + 1)^{-1} < b^{-1}$, Theorem 1.1 implies that $\|b\theta\| < b^{-1}$. Furthermore, this inequality is satisfied by infinitely many positive integers b for the following reason. Suppose there were only a finite number of such integers, say $b_1, b_2, \cdots, b_r$ with

$$\|b_j\theta\| < b_j^{-1} \qquad \text{for } j = 1, 2, \cdots, r.$$

Then choose the integer m so large that

$$\frac{1}{m} < \|b_j\theta\|$$

holds for every $j = 1, 2, \cdots, r$. Then apply Theorem 1.1 with this value of m, and note that this process yields an integer b such that

$$\|b\theta\| < \frac{1}{m+1} < \frac{1}{m} < \|b_j\theta\|, \qquad j = 1, 2, \cdots, r.$$

Hence b is different from each of $b_1, b_2, \cdots, b_r$. Also $\|b\theta\| < b^{-1}$, so there can be no end to the integers satisfying this inequality. The following corollary states what we have just proved.

COROLLARY 1.2. *Given any irrational number θ, there are infinitely many rational numbers a/b, where a and $b > 0$ are integers, such that*

$$\left|\theta - \frac{a}{b}\right| < \frac{1}{b^2}. \tag{4}$$

Note that this result is a considerable improvement over inequality (1). It is natural to ask whether Corollary 1.2 can also be improved, for instance, by the replacement of $1/b^2$ by $1/b^3$. It cannot; in fact, Corollary 1.2 becomes false if $1/b^2$ is replaced by $1/b^{2+\epsilon}$ for any positive ϵ. Nevertheless, although the exponent cannot be improved, this corollary can be strengthened by a constant factor in (4). Specifically $1/b^2$ can be replaced by $1/(\sqrt{5}b^2)$, and no larger constant can be used than $\sqrt{5}$. This result, due to Hurwitz, is proved in the next section.

1.2. *The theorem of Hurwitz*

We first prove a preliminary result about Farey sequences. For any positive integer n, the Farey sequence F_n is the sequence, ordered in size, of all rational fractions a/b in lowest terms with $0 < b \leq n$. For example,

$$F_7\colon \quad \cdots, \frac{-1}{6}, \frac{-1}{7}, \frac{0}{1}, \frac{1}{7}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{2}{7}, \frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \cdots.$$

Of the many known properties of Farey sequences, only two are needed for our purposes, as follows.

THEOREM 1.3. *If a/b and c/d are two consecutive terms in F_n, then, presuming a/b to be the smaller, $bc - ad = 1$. Furthermore, if θ is any* *given irrational number, and if r is any positive integer, then for all n sufficiently large the two fractions a/b and c/d adjacent to θ in F_n have denominators larger than r, that is, $b > r$ and $d > r$.*

Proof: The proof of the first part is by induction on n. If $n = 1$, then $b = 1$, $d = 1$, and $c = a + 1$, so that

$$bc - ad = a + 1 - a = 1.$$

Next we suppose that the result holds for F_n, and prove it for F_{n+1}. Let a/b and c/d be adjacent fractions in F_n. First we note that $b + d \geqq n + 1$, since otherwise the fraction $(a + c)/(b + d)$, reduced if necessary, would belong to F_n. But this is not possible since

$$\frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}.$$

Now with respect to F_{n+1} there are two possibilities: first that a/b and c/d are adjacent, and the second that some fraction or fractions lie between. In the first case there is nothing to prove because $bc - ad = 1$ by the induction hypothesis. In the second case, any such fraction, being in F_{n+1} but not in F_n, has denominator $n + 1$. Denoting the fraction by $k/(n + 1)$, we write

$$\frac{1}{bd} = \frac{c}{d} - \frac{a}{b} = \frac{c}{d} + \frac{k}{n+1} + \frac{k}{n+1} - \frac{a}{b}$$

$$= \frac{u}{d(n+1)} + \frac{v}{b(n+1)},$$

where $u = c(n + 1) - dk \geqq 1$, $v = bk - a(n + 1) \geqq 1$. Our aim is to establish that $u = 1$ and $v = 1$. If on the contrary $u > 1$ or $v > 1$ or both, then it follows that

$$\frac{1}{bd} > \frac{1}{d(n+1)} + \frac{1}{b(n+1)}, \qquad n + 1 > b + d,$$

which is contrary to what was established earlier. Hence $u = 1$ and $v = 1$, and so

$$\frac{c}{d} - \frac{k}{n+1} = \frac{1}{d(n+1)}, \qquad \frac{k}{n+1} - \frac{a}{b} = \frac{1}{b(n+1)}.$$

Finally, we observe that at most one fraction can occur between a/b and c/d in F_{n+1}. For if there were another fraction besides $k/(n+1)$ it must have the form $h/(n+1)$. Then the preceding argument implies that

$$\frac{h}{n+1} - \frac{a}{b} = \frac{1}{b(n+1)},$$

and hence

$$\frac{h}{n+1} - \frac{a}{b} = \frac{k}{n+1} - \frac{a}{b}, \qquad h = k.$$

This completes the proof of the first part of Theorem 1.3.

To prove the second part, let $m_1, m_2, \cdots, m_r$ denote the integers nearest to $\theta, 2\theta, \cdots, r\theta$. Choose n sufficiently large so that for every $j = 1, 2, \cdots, r$,

$$\frac{1}{n} < \left|\theta - \frac{m_j}{j}\right|.$$

If q is any integer, then for every $j = 1, 2, \cdots, r$,

$$|j\theta - m_j| \leqq |j\theta - q|, \qquad \left|\theta - \frac{m_j}{j}\right| \leqq \left|\theta - \frac{q}{j}\right|, \qquad \frac{1}{n} < \left|\theta - \frac{q}{j}\right|.$$

Now the difference between adjacent fractions in F_n does not exceed $1/n$, because F_n contains all fractions with denominator n, of which some are perhaps in reduced form. Hence if a/b and c/d are the fractions adjacent to θ in F_n, we see that

$$\left|\theta - \frac{a}{b}\right| < \left|\frac{c}{d} - \frac{a}{b}\right| \leqq \frac{1}{n},$$

and

$$\left|\theta - \frac{c}{d}\right| < \left|\frac{c}{d} - \frac{a}{b}\right| \leqq \frac{1}{n}.$$

A comparison of these with the previous inequalities establishes that $b > r$ and $d > r$, and the proof of Theorem 1.3 is complete.

Another result we shall need is the following.

LEMMA 1.4. *There are no positive integers x and y which satisfy simultaneously the inequalities*

$$(5) \quad \frac{1}{xy} \geqq \frac{1}{\sqrt{5}}\left(\frac{1}{x^2}+\frac{1}{y^2}\right) \quad \text{and} \quad \frac{1}{x(x+y)} \geqq \frac{1}{\sqrt{5}}\left(\frac{1}{x^2} + \frac{1}{(x+y)^2}\right).$$

Proof: If there were such integers, then from (5) it would follow that

$$0 \geqq x^2 + y^2 - \sqrt{5}xy \quad \text{and} \quad 0 \geqq (2 - \sqrt{5})(x^2 + xy) + y^2.$$

Adding these, we get

$$0 \geqq \tfrac{1}{2}\{(\sqrt{5} - 1)x - 2y\}^2$$

which is false for rational x/y.

We are now in a position to prove a basic result (Hurwitz, 1891).

THEOREM 1.5. *Given any irrational number θ there exist infinitely many rational numbers h/k in lowest terms such that*

$$\left|\theta - \frac{h}{k}\right| < \frac{1}{\sqrt{5}k^2}. \tag{6}$$

Furthermore, this inequality is best possible in the sense that the result becomes false if $\sqrt{5}$ is replaced by any larger constant.

Proof: Locate θ between two consecutive fractions of the Farey sequence F_n, say $a/b < \theta < c/d$ with b and d positive. We consider two cases according to whether θ is greater or less than $(a + c)/(b + d)$. In case $\theta > (a + c)/(b + d)$, we prove that not all three of the inequalities

$$\theta - \frac{a}{b} \geqq \frac{1}{\sqrt{5}b^2}, \qquad \theta - \frac{a + c}{b + d} \geqq \frac{1}{\sqrt{5}(b + d)^2}, \qquad \frac{c}{d} - \theta \geqq \frac{1}{\sqrt{5}d^2}$$

can hold. For if we add the first and third of these, and then the second and third, we get (5) with $x = d$ and $y = b$.

In the other case, $\theta < (a + c)/(b + d)$, we prove that not all three of the inequalities

$$\theta - \frac{a}{b} \geqq \frac{1}{\sqrt{5}b^2}, \qquad \frac{a + c}{b + d} - \theta \geqq \frac{1}{\sqrt{5}(b + d)^2}, \qquad \frac{c}{d} - \theta \geqq \frac{1}{\sqrt{5}d^2}$$

can hold. For if we add the first and third of these, and then the first and second, we get (5) with $x = b$ and $y = d$.

Hence the inequality (6) holds with h/k replaced by at least one of a/b, c/d, and $(a + c)/(b + d)$. To prove that there are infinitely many solutions of (6) we argue as follows. Suppose there were only a finite number of solutions h/k, and we let r denote the maximum denominator among these solutions. Then the second part of Theorem 1.3 guarantees that for sufficiently large n the consecutive fractions a/b and c/d adjacent to θ in F_n have denominators greater than r. This process then gives a solution h/k to (6) of one of the three forms a/b, c/d, or $(a + c)/(b + d)$. Now a/b and c/d are in lowest terms by definition of Farey sequences. Also $(a + c)/b + d)$ is in lowest terms because

$$c(b + d) - d(a + c) = bc - ad = 1,$$

so that any common divisor of $a + c$ and $b + d$ is a divisor of 1. Thus the solution of (6) so obtained is in lowest terms and its denominator exceeds those of previously obtained solutions.

To complete the proof of the theorem we must show that $\sqrt{5}$ is the best possible constant. Before doing that, we remark on the proof thus far. The proof that (6) has infinitely many solutions has a slightly artificial aspect in that the inequalities (5) seem to have no motivating source. It might appear that some variation on the inequalities (5) would lead to better results. This is not the case, as we now show by establishing that the constant $\sqrt{5}$ in (6) is best possible.

Let θ_0 and θ_1 be defined by

$$\theta_0 = \frac{1 + \sqrt{5}}{2}, \qquad \theta_1 = \frac{1 - \sqrt{5}}{2},$$

so that

$$(x - \theta_0)(x - \theta_1) = x^2 - x - 1.$$

For any integers h and k, with $k > 0$, we see that

$$\left|\frac{h}{k} - \theta_0\right| \cdot \left|\frac{h}{k} - \theta_1\right| = \left|\left(\frac{h}{k}\right)^2 - \frac{h}{k} - 1\right| \neq 0.$$

Also $\theta_1 = \theta_0 - \sqrt{5}$ and so

$$\left|\frac{h}{k} - \theta_0\right| \cdot \left|\frac{h}{k} - \theta_1 + \sqrt{5}\right| = \frac{|h^2 - hk - k^2|}{k^2} \geq \frac{1}{k^2}.$$

An application of the triangle inequality gives

$$\frac{1}{k^2} \leq \left|\frac{h}{k} - \theta_0\right| \cdot \left\{\left|\frac{h}{k} - \theta_0\right| + \sqrt{5}\right\}. \tag{7}$$

Now if for some positive number β there are infinitely many h_j/k_j, $j = 1, 2, 3, \cdots$, such that

$$\left|\frac{h_j}{k_j} - \theta_0\right| < \frac{1}{\beta k_j^2}, \tag{7a}$$

then $k_j \to \infty$ as $j \to \infty$. Furthermore, from (7) we get

$$\frac{1}{k_j^2} < \frac{1}{\beta k_j^2}\left(\frac{1}{\beta k_j^2} + \sqrt{5}\right),$$

$$\beta < \frac{1}{\beta k_j^2} + \sqrt{5},$$

$$\beta \leqq \lim_{j\to\infty}\left(\frac{1}{\beta k_j^2} + \sqrt{5}\right) = \sqrt{5}.$$

Hence $\sqrt{5}$ is the largest possible constant in (6).

Thus the theorem is proved, and we note that the exponent 2 on the k^2 in (6) is best possible. That is, if γ is any fixed real number >2, and c is any positive constant, there are only finitely many h/k satisfying

$$\left|\theta_0 - \frac{h}{k}\right| < \frac{1}{ck^\gamma}.$$

For if not, we could obtain infinitely many h/k satisfying (7*a*) with, say $\beta = 3$.

Next we formulate a simple consequence of Theorem 1.5.

COROLLARY 1.6. *Given any real numbers a_1, a_2, b_1, b_2 with $\Delta \neq 0$, where $\Delta = |a_1b_2 - a_2b_1|$, and given any positive ε, there are infinitely many pairs of integers h, k such that*

$$|a_1k + b_1h| \cdot |a_2k + b_2h| < \frac{\Delta}{\sqrt{5}} + \varepsilon.$$

Proof: If any one of a_1, a_2, b_1, b_2 is zero, or if a_1/b_1 or a_2/b_2 is rational, the result is immediate. So suppose that a_1/b_1 is irrational. Then let $-a_1/b_1$ play the role of θ in Theorem 1.5 and let h/k be any one of the rational numbers in that result. Define δ by

$$\frac{h}{k} + \frac{a_1}{b_1} = \frac{\delta}{k^2}$$