Continued Fractions in Approximation and Number Theory

CONTINUED FRACTIONS

IN

RATIONAL APPROXIMATIONS

AND

NUMBER THEORY

by

DAVID C. EDWARDS

Thesis for Master of Science
Department of Mathematics
McGill University
January 1971

TABLE OF CONTENTS

INTRODUCTION

This thesis surveys the elementary theory of continued fractions and discusses in detail some important applications of continued fractions to the theory of rational approximations to real numbers and elementary number theory. Chapters 1 to 5 introduce continued fractions and their basic properties, and provide the results on which the following chapters are based. Chapters 6 and 7 develop the elementary theory of rational approximations, culminating with Hurwitz's Theorem, and using continued fractions as the fundamental tool. The approach is essentially that of Khintchine [4], Sections I and II, but the theorems are stated and proved in greater detail, an attempt is made to clearly motivate the definitions, and some closely related results from [6], Chapter 7, are included. Chapters 8 to 12 investigate closely the applications of continued fractions to the Euclidean Algorithm and to the Pell Equation $x^2-Dy^2= N$. This involves a thorough examination of periodic continued fractions, in particular the simple-continued-fraction expansion of $\sqrt{D}$ (D being a positive nonsquare integer). The material is drawn primarily from Perron [8], Olds [7], and Niven and Zuckerman [6]. As far as I have been able to discover, the bound on the period given in Theorems 9.4 and 10.1 is an original result, although admittedly a minor one. Theorems 12.2 to 12.4 (from Perron [8]) are significant and fairly deep results, rarely found in discussions of continued fractions or the Pell Equation.

Most recent texts on number theory treat the theory of continued fractions fleetingly, or with a single application in mind ; in contrast, an attempt has been made in this survey to demonstrate the richness and wide applicability of the theory.

PART I

SOME BASIC RESULTS

Chapter 1. DEFINITIONS, NOTATIONS, AND THE BASIC FORMULAS.

Let F be a field. A continued fraction in F is a sequence $(a_1, b_2, a_2, b_3, a_3, \ldots)$ of an odd or infinite number of elements of F such that, for each $n \geq 2$, each denominator in the following composite fraction is not zero :

$$a_1 + \cfrac{b_2}{a_2 + \cfrac{b_3}{a_3 + \cfrac{\ddots}{\ddots + \cfrac{b_{n-1}}{a_{n-1} + \cfrac{b_n}{a_n}}}}} \qquad (1)$$

The fraction (1) is called the nth convergent of the continued fraction, and denoted by c_n. It is usually written in the more convenient form

$$c_n = a_1 + \frac{b_2}{a_2} + \frac{b_3}{a_3} + \cdots + \frac{b_n}{a_n} . \qquad (2)$$

Additionally, the first convergent is $c_1 = a_1$.

If the sequence is finite, say $(a_1, b_2, a_2, \ldots, b_m, a_m)$ where $m \geq 1$, then the continued fraction is called a finite continued fraction, and its value is defined to be the last convergent, c_m. If the sequence is infinite, then we have an infinite sequence of convergents $c_1, c_2, c_3 \ldots$, and (assuming a metric on F) the continued fraction is said to be convergent or divergent in F according as $\lim_{n\to\infty} c_n$ exists or not in F. In the former case, the limit is taken as the value of the continued

fraction. Normally F is the real field.

When $b_2, b_3, \ldots$ are all 1, the continued fraction is said to have unit numerators, and we use the notation

$$(a_1, 1, a_2, 1, a_3, \ldots) = [a_1, a_2, a_3, \ldots] . \tag{3}$$

For a finite or convergent continued fraction, we often use the same notation for the continued fraction and its value. Thus we write $c_n = [a_1, a_2, \ldots, a_n]$ for the convergents of a continued fraction with unit numerators.

A continued fraction $[a_1, a_2, \ldots]$, with unit numerators and infinitely many terms, is said to be <u>periodic</u> with period t if there exist $m \geq 0$ and $t \geq 1$ such that for all $n > m$, $a_{n+t} = a_n$. We write

$$[a_1, a_2, \ldots] = [a_1, \ldots, a_m, \overline{a_{m+1}, \ldots, a_{m+t}}] . \tag{4}$$

The least period t_o is called the <u>fundamental period</u>. t_o divides any period t, for if not, let $t = k\, t_o + r$, where $0 < r < t_o$; then for any $n > m$, we have $a_n = a_{n+t} = a_{n+k\, t_o + r} = a_{n+r}$, therefore r is a period, contradicting that t_o is the least period. The continued fraction is said to be <u>purely periodic</u> if it is possible to take $m = 0$, i.e.

$$[a_1, a_2, \ldots] = [\overline{a_1, a_2, \ldots, a_t}] . \tag{5}$$

Given a continued fraction, it is desirable to have a systematic method for expressing each convergent c_n as a simple fraction p_n/q_n, as in $c_1 = a_1/1$, $c_2 = (a_1 a_2 + b_2) / a_2$, etc. The key to finding such a method is to consider c_n as a function of a_n. Now if we simplify (1) starting at the lower right and working upwards, we find

That C_n is a quotient of linear functions of a_n. In fact, suppose $c_n = p_n/q_n$, where $p_n = k_n a_n + \ell_n$ and $q_n = r_n a_n + s_n$, and k_n, ℓ_n, r_n and s_n are independent of a_n (i.e. k_n, ℓ_n, r_n, s_n depend only on $a_1, b_2, a_2, \ldots a_{n-1}, b_n$).

Then $C_{n+1} = (a_1, b_2, a_2, \ldots, b_n, a_n + b_{n+1}/a_{n+1})$

$$= \frac{k_n (a_n + b_{n+1}/a_{n+1}) + \ell_n}{r_n (a_n + b_{n+1}/a_{n+1}) + s_n}$$

$$= \frac{(k_n a_n + \ell_n) a_{n+1} + k_n b_{n+1}}{(r_n a_n + s_n) a_{n+1} + r_n b_{n+1}}$$

$$= \frac{p_n a_{n+1} + k_n b_{n+1}}{q_n a_{n+1} + r_n b_{n+1}} \quad .$$

Thus we can take $p_{n+1} = p_n a_{n+1} + k_n b_{n+1}$ and $q_{n+1} = q_n a_{n+1} + r_n b_{n+1}$, and c_{n+1} is also a quotient of linear functions of a_{n+1}. Applying the same argument to c_{n+1}, we therefore have $c_{n+2} = p_{n+2}/q_{n+2}$, where $p_{n+2} = p_{n+1} a_{n+2} + p_n b_{n+2}$ and $q_{n+2} = q_{n+1} a_{n+2} + q_n b_{n+2}$.

Noting that $c_1 = (1.a_1 + 0) / (0.a_1 + 1)$ and using induction, we clearly have the following theorem :

Theorem 1.1 The convergents $c_1, c_2, \ldots$ of the continued fraction $(a_1, b_2, a_2, b_3, a_3, \ldots)$ are $c_n = p_n/q_n$, where p_n and q_n are defined as follows :

$$p_1 = a_1 \ , \ q_1 = 1$$

$$p_2 = a_1 a_2 + b_2 \ , \ q_2 = a_2$$

$$p_n = a_n p_{n-1} + b_n p_{n-2} \qquad (n \geq 3) \qquad (6)$$

$$q_n = a_n q_{n-1} + b_n q_{n-2} \qquad (n \geq 3) \qquad (7)$$

Theorem 1.2. Following the notation of the previous theorem,

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^n \prod_{i=2}^{n} b_i \qquad (n \geq 2) \qquad (8)$$

$$c_n - c_{n-1} = \frac{(-1)^n \prod_{i=2}^{n} b_i}{q_n q_{n-1}} \qquad (n \geq 2) \qquad (9)$$

$$c_n - c_{n-2} = \frac{a_n (-1)^{n-1} \prod_{i=2}^{n} b_i}{q_n q_{n-2}} \qquad (n \geq 3) \qquad (10)$$

Proof: $p_2 q_1 - p_1 q_2 = a_1 a_2 + b_2 - a_1 a_2 = b_2$, therefore (8) is true for $n = 2$. Assume (8) for $n-1$. Multiplying (6) by q_{n-1} and (7) by p_{n-1} and subtracting, $p_n q_{n-1} - p_{n-1} q_n = -b_n (p_{n-1} q_{n-2} - p_{n-2} q_{n-1}) = (-1)^n \prod_{i=2}^{n} b_i$, and (8) is proved.

(9) follows from (8) by dividing by $q_n q_{n-1}$. Using (9), $c_n - c_{n-2} = (c_n - c_{n-1}) + (c_{n-1} - c_{n-2}) = \frac{(-1)^{n-1} \prod_{i=2}^{n-1} b_i}{q_{n-1}} \left(\frac{-b_n}{q_n} + \frac{1}{q_{n-2}} \right) =$

$\frac{a_n (-1)^{n-1} \prod_{i=2}^{n-1} b_i}{q_n q_{n-2}}$, since by (7), $-b_n q_{n+2} + q_n = a_n q_{n-1}$. //

Chapter 2. CONVERGENCE OF CONTINUED FRACTIONS WITH UNIT NUMERATORS.

From now on, all continued fractions will be assumed to have unit numerators. For the present chapter at least, this restriction is not serious, because it is clear that any continued fraction with nonzero numerators can be converted to an equivalent (i.e. same convergents) continued fraction with unit numerators. Also, for definiteness, we shall work in the real field.

Introducing $q_0 = 0 = p_{-1}$ and $q_{-1} = 1 = p_0$, it may be checked that formulas (6) and (7) of Chapter 1 are also valid for n=1 and 2, and that (8) is valid for n = 0 and 1, assuming $b_i = 1$ for all i. For the continued fraction

$$[a_1, a_2, a_3, \ldots] \tag{1}$$

(where $a_1, a_2, a_3, \ldots$ are any real numbers), the formulas of Chapter 1 therefore become :

$$p_n = a_n p_{n-1} + p_{n-2} \qquad (n \geq 1) \qquad (2)$$

$$q_n = a_n q_{n-1} + q_{n-2} \qquad (n \geq 1) \qquad (3)$$

$$c_n = [a_1, a_2, \ldots, a_n] = p_n/q_n \qquad (n \geq 1) \qquad (4)$$

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^n \qquad (n \geq 0) \qquad (5)$$

$$c_n - c_{n-1} = \frac{(-1)^n}{q_n q_{n-1}} \qquad (n \geq 2) \qquad (6)$$

$$c_n - c_{n-2} = \frac{a_n (-1)^{n-1}}{q_n q_{n-2}} \qquad (n \geq 3) \qquad (7)$$

Remark: If $a_2, a_3, \ldots$ are positive, then (1) is a valid continued fraction (since all denominators are positive), although not necessarily

convergent, and induction on (3) shows that $q_1, q_2, \ldots$ are positive.

Theorem 2.1 Let $a_2, a_3, \ldots$ be positive real numbers. Then the odd-numbered convergents of (1) form a strictly increasing sequence, the even-numbered convergents form a strictly decreasing sequence, and every odd-numbered convergent is less than every even-numbered one ; (1) is convergent if and only if

$$\lim_{n \to \infty} q_n q_{n-1} = \infty . \qquad (8)$$

Proof: The first three assertions follow immediately from formula (6) and (7). It is then clear that (1) is convergent if and only if $c_{2n} - c_{2n-1} \to 0$ (equivalently, $c_{2n+1} - c_{2n} \to 0$), which, by (6), is equivalent to (8). //

Corollary: If $x = [a_1, a_2, \ldots]$, where $a_2, a_3, \ldots > 0$, then x lies strictly between any two consecutive convergents (except the last two if the continued fraction is finite).

The following important theorem provides a convenient necessary and sufficient condition for convergence.

Theorem 2.2 Let $a_2, a_3, \ldots$ be positive real numbers. Then (1) converges if and only if the series

$$\sum_{n=1}^{\infty} a_n \qquad (9)$$

is divergent.

Proof: The proof uses the well known result that if $0 < t_n < 1$, then the $\prod_{n=1}^{\infty} (1-t_n)$ is convergent (i.e. has a positive limit) if and only if

$\sum_{n=1}^{\infty} t_n$ is convergent. Suppose (9) converges. By (3), $q_n > q_{n-2}$, hence (a) $q_{n-1} < q_n$ or (b) $q_{n-1} > q_{n-2}$. $a_n \to 0$, therefore there exists N such that $a_n < 1$ if $n \geq N$. For $n \geq N$, (3) gives $q_n < a_n q_n + q_{n-2}$, i.e. $q_n < q_{n-2}/(1-a_n)$, in case (a), and $q_n < a_n q_{n-1} + q_{n-1} < q_{n-1}/(1-a_n)$ in case (b). Therefore

$$q_n < \frac{q_r}{1-a_n} \quad (r = n-1 \text{ or } n-2) .$$

Repeated application of this result gives $r,\ldots,s,t$ such that $n > r > \ldots > A \geq N$, $t = N-1$ or $N-2$, and

$$q_n < \frac{q_t}{(1-a_n)(1-a_r)\ldots(1-a_s)} \tag{10}$$

Now $\prod_{i=N}^{\infty} (1-a_i) = L > 0$; the denominator of (10) exceeds L, therefore letting M be the larger of q_{N-1} and q_{N-2}, we have $q_n < M/L$ $(n \geq N)$, hence $q_{n+1} q_n < M^2/L^2$, so that (1) diverges, by Theorem 2.1.

Conversely, suppose (9) diverges. Now (3) gives $q_n > q_{n-2} > \ldots > q_2$ (n even) and $q_n > q_{n-2} > \ldots > q_1$ (n odd). Let $c = \max(q_1, q_2)$. Then (3) gives $q_n \geq q_{n-2} + c a_n$ $(n \geq 2)$. Therefore :

$$\begin{aligned} q_n + q_{n-1} &\geq q_{n-2} + q_{n-3} + c\,(a_n + a_{n-1}) \\ &\geq q_{n-4} + q_{n-5} + c\,(a_n + a_{n-1} + a_{n-2} + a_{n-3}) \\ &\geq \cdots \\ &\geq \begin{cases} q_2 + q_1 + c \sum_{i=3}^{n} a_i & (n \text{ even}) \\ q_1 + q_o + c \sum_{i=2}^{n} a_i & (n \text{ odd}) \end{cases} \\ &> c\, S_n \text{ , where } S_n = \sum_{2=3}^{n} a_i . \end{aligned}$$

Therefore at least one of q_n, q_{n-1} exceeds $\frac{1}{2} c S_n$. The other is at least c, hence $q_n q_{n-1} > \frac{c^2}{2} S_n$. By assumption, $S_n \to \infty$, therefore (1) converges, by Theorem 2.1. //

Theorem 2.3. Let $a_2, a_3, \ldots$ be positive real numbers and let $n \geq 1$. The continued fraction

$$[a_n, a_{n+1}, \ldots] \tag{11}$$

is convergent if and only if (1) is convergent. If (1) and (11) converge to values x and m_n respectively, then

$$x = [a_1, a_2, \ldots, a_{n-1}, m_n] . \tag{12}$$

Proof: (By theorem 2.1, $M_n > a_n > 0$ for $n \geq 2$, hence (12) is a continued fraction.) Let $c_r = [a_1, a_2, \ldots, a_r]$.

The result is trivial for $n = 1$. For $n = 2$,

$$c_r = a_1 + \frac{1}{[a_2, \ldots, a_r]} \tag{13}$$

or

$$[a_2, \ldots, a_r] = \frac{1}{c_r - a_1} . \tag{14}$$

Now if (11) converges to M_2, then $M_2 \neq 0$ as noted, and taking limits in (13) shows that (1) converges to $a_1 + 1/m_2 = [a_1, m_2]$; if (1) converges to x, then ($x - a_1 \neq 0$ by Theorem 2.1) taking limits in (14) shows that (11) converges, and again (12) holds. For $n \geq 3$, (1) converges if and only if $[a_2, a_3, \ldots]$ does, and, assuming the theorem for $n-1$, the latter converges if and only if (11) converges; also, using the theorem for 2 and $n-1$, $x = [a_1, [a_2, a_3, \ldots]] = [a_1, [a_2, \ldots, a_{n-1}, m_n]] = [a_1, \ldots, a_{n-1}, m_n]$. //

Chapter 3. SIMPLE CONTINUED FRACTIONS.

A simple continued fraction is one of the form $[a_1,a_2,a_3,\ldots]$, where a_1 is any integer and $a_2,a_3,\ldots$ are positive integers. In the following chapters we shall be concerned primarily with this special type of continued fraction.

For a simple continued fraction, the numbers p_n, q_n are integers, and furthermore it is clear from (3) of Chapter 2 that $1=q_1 \leq q_2 < q_3 <\ldots$. In particular, $\lim_{n\to\infty} q_n q_{n-1} = \infty$, therefore by Theorem 2.1 any simple continued fraction is convergent ; this also follows from Theorem 2.2, since $\sum_{n=1}^{\infty} a_n = \infty$.

Theorem 3.1 The integers p_n, q_n are relatively prime for any n .

Proof: By (5) of Chapter 2, any common divisor of p_n and q_n divides $(-1)^n$, therefore $(p_n, q_n) = 1$. //

Theorem 3.2 Let x be any real number. Then x is the value of the (finite or infinite) simple continued fraction $[a_1, a_2,\ldots]$, where a_i are defined inductively as follows ($[x]$ denotes the integral part of x) :

$$m_1 = x \quad , \quad a_1 = [x]$$

$$m_{i+1} = \frac{1}{m_i - a_i} \quad , \quad a_{i+1} = [m_{i+1}] \ . \qquad (1)$$

The induction terminates with a_n if m_n is found to be an integer; then $x = [a_1,a_2,\ldots,a_n]$ and , if $n\neq 1$, we have $a_n>1$. If no m_i is an integer, the continued fraction is infinite.

Proof: It is clear that the process can be continued as long as m_i remains nonintegral, and yields integers $a_1, a_2,\ldots$ with $a_2,a_3,\ldots$positive.

Furthermore, if $m_i (i \neq 1)$ is an integer then $a_i = m_i > 1$, since $M_{i-1} - a_{i-1} < 1$. It remains to show that $x = [a_1, a_2, \ldots]$. Now $m_i = a_i + 1/m_{i+1}$, therefore it is clear by induction that $x = [a_1, a_2, \ldots, a_i, m_{i+1}]$ $(i \geq 0)$. If the process (1) terminates with $m_n = a_n$, this proves $x = [a_1, a_2, \ldots, a_n]$. If it does not terminate, formula (6) of Chapter 2 applied to $[a_1, \ldots, a_i, m_{i+1}]$ gives $| x - c_i | = \dfrac{1}{q_i q'_{i+1}}$, where the prime refers to $[a_1, \ldots, a_i, m_{i+1}]$. Hence $| x - c_i | < 1/q_i \to 0$, therefore $x = [a_1, a_2, \ldots]$. //

Theorem 3.3 The value of any finite simple continued fraction is rational. Conversely, if $x = r/s$ $(r, s$ integers, $s > 0)$, then the procedure in Theorem 3.2 gives $x = [a_1, a_2, \ldots, a_n]$, where $a_1, \ldots, a_n$ are the quotients in the Euclidean algorithm for r and s :

$$
\begin{aligned}
r &= a_1 s + r_1 && (0 < r_1 < s) \\
s &= a_2 r_1 + r_2 && (0 < r_2 < r_1) \\
&\cdot \\
&\cdot \\
&\cdot \\
r_{n-3} &= a_{n-1} r_{n-2} + r_{n-1} && (0 < r_{n-1} < r_{n-2}) \\
r_{n-2} &= a_n r_{n-1} .
\end{aligned}
\qquad (2)
$$

Proof: The first assertion is obvious. For the second, divide the equations of (2) by $s, r_1, \ldots, r_{n-2}, r_{n-1}$ respectively, and denote the resulting left hand side, by $m_1, m_2, \ldots, m_n$. Clearly m_i and a_i are the same as in Theorem 3.2. //

Theorem 3.4 The representation of any irrational number as a simple

continued fraction is unique. Any rational number is the value of exactly two simple continued fractions, and these are of the form $[a_1, a_2,\ldots,a_n]$ ($a_n>1$ if $n \neq 1$) and $[a_1,\ldots, a_n-1, 1]$.

Proof: Let x be irrational and suppose $x = [a_1, a_2,\ldots]$. This continued fraction must be infinite, for otherwise x would be rational. To prove uniqueness, it is sufficient to prive that a_i are the same integers as given by the procedure of Theorem 3.2. But this is clear, because letting $m_i = [a_i, a_{i+1},\ldots]$ we have $m_1 = x$, $m_i = a_i+1/m_{i+1}$, and $m_{i+1} > 1$. Now let x be rational and suppose $x = [b_1, b_2,\ldots]$ (finite or infinite). Let $x = [a_1, a_2,\ldots, a_n]$ ($a_n > 1$ if $n \neq 1$), e.g. the expansion given by Theorems 3.2 and 3.3. We shall prove by induction on n that $[b_1,b_2,\ldots]$ is $[a_1,a_2,\ldots,a_n]$ or $[a_1,a_2,\ldots,a_{n-1},1]$. Let m (possibly infinite) be the number of terms in $[b_1,b_2,\ldots]$.

For $n = 1$, $x = a_1$ is an integer, and we consider 3 cases :

(i) $m = 1$: $b_1 = x = a_1$ as required.

(ii) $m = 2$: $x = b_1 + 1/b_2$, therefore $b_2 = 1$, and

$[b_1, b_2] = [a_1 - 1 , 1]$ as required .

(iii) $m > 2$: $b_1 < x < b_1 + 1/b_2$, contradicting that x is an integer. Therefore this case does not occur. For $n>1$, x is not an integer and we have $m \geq 2$, $b_1 < x \leq b_1 + 1/b_2 \leq b_1 + 1$, therefore $b_1=[x] = a_1$, and $1/(x-b_1) = [b_2, b_3,\ldots] = [a_2,\ldots,a_n]$; now the required result follows by applying the result for $n-1$. //

Corollary: The value of any infinite simple continued fraction is irrational.

Proof: The value cannot be rational, since the only expansions of a

rational number are the two finite ones mentioned in the theorem. //

Chapter 4. THE FIBONACCI NUMBERS.

In this chapter we consider some elementary properties of the Fibonacci numbers and the closely related simple continued fraction $[1,1,1,\ldots]$. This continued fraction plays a special role in later chapters.

Let t denote the value of $[1,1,1,\ldots]$. Clearly $t=1+1/[1,1,\ldots] = 1+1/t$, i.e.

$$t^2 - t - 1 = 0 . \tag{1}$$

Solving (1) for the positive root, we find

$$t = \frac{1+\sqrt{5}}{2} = 1.618... \tag{2}$$

the other root is

$$s = -\frac{1}{t} = \frac{1-\sqrt{5}}{2} = -0.618... \tag{3}$$

Formulas (2) and (3) of Chapter 2, applied to $[1, 1, 1,\ldots]$, give $p_n = p_{n-1} + p_{n-2}$, $q_n = q_{n-1} + q_{n-2}$. Furthermore $p_o = p_1 = 1$ and $q_1 = q_2 = 1$. Now the Fibonacci numbers $F_1, F_2, F_3,\ldots$ are defined by :

$$\begin{aligned} F_1 &= F_2 = 1 \\ F_n &= F_{n-1} + F_{n-2} . \end{aligned} \tag{4}$$

Therefore we have

$$\begin{aligned} p_n &= F_{n+1} \\ q_n &= F_n , \end{aligned} \tag{5}$$

and the convergents of $[1, 1, 1,\ldots]$ are ratios of consecutive Fibonacci

numbers, namely $c_n = F_{n+1}/F_n$.

It is important to know how quickly the Fibonacci numbers grow with increasing n . From (1) we note that $t^n = t^{n-1} + t^{n-2}$, and the same for s. Therefore for any a and b, $H_n = at^n + bs^n$ satisfies $H_n = H_{n-1} + H_{n-2}$; solving for a and b which result in $H_o = 0$, $H_1 = 1$ (i.e. $H_o = F_o$, $H_1 = F_1$) , we get $a = -b = 1/\sqrt{5}$. Therefore $H_n = F_n$ and we have proved the very interesting formula (called Binet's Formula) :

$$F_n = \frac{1}{\sqrt{5}} \left(\left(\frac{1 + \sqrt{5}}{2} \right)^n - \left(\frac{1 - \sqrt{5}}{2} \right)^n \right) . \tag{6}$$

It follows that

$$\lim_{n \to \infty} \frac{F_n}{t^n} = \frac{1}{\sqrt{5}} , \tag{7}$$

so that $F_1, F_2, \ldots$ approximates a geometric progression. In fact, it is clear from (6) that F_n is the nearest integer to $t^n/\sqrt{5}$.

The following theorem shows that the denominators of the convergents of any simple continued fraction increase at least as fast as the Fibonacci numbers.

Theorem 4.1. For any simple continued fraction, $q_n \geq F_n$.

Proof: $q_1 = 1 = F_1$ and $q_2 = a_2 \geq 1 = F_2$. Assuming $q_{n-2} \geq F_{n-2}$ and $q_{n-1} \geq F_{n-1}$, we have $q_n = a_n q_{n-1} + q_{n-2} \geq q_{n-1} + q_{n-2} \geq F_{n-1} + F_{n-2} = F_n$. //

Chapter 5. MISCELLANEOUS RESULTS.

Here we mention some miscellaneous theorems which will be used in later chapters, but which either are not of general interest or are well known elementary results.

Theorem 5.1. Let $x > 1$ and consider the simple continued fraction expansions of x and $1/x$. For $n \geq 2$, the nth convergent of $1/x$ is the reciprocal of the $(n-1)$ st convergent of x.

Proof: If $x = [a_1, a_2, \ldots]$, then $1/x = [0, a_1, a_2, \ldots]$. Letting primed quantities refer to $1/x$, and using (2) and (3) of Chapter 2, one can verify by induction that $p'_n = q_{n-1}$, $q'_n = p_{n-1}$, from which the result follows. Details may be found in [6],

Theorem 7.15. //

Theorem 5.2. Let $a_2, a_3, \ldots$ be positive real numbers and assume that $x = [a_1, a_2, \ldots] \neq 0$. Denote $[a_n, a_{n+1}, \ldots]$ by m_n. Then for $n \geq 1$,

$$\frac{1}{m_{n+1}} = [a_n, a_{n-1}, \ldots, a_1, -\frac{1}{x}]. \tag{1}$$

Proof: (Note : in the course of the proof, it must be shown that (1) is indeed a valid continued fraction, i.e. has no zero denominators.) For $n = 1$, $[a_1, -\frac{1}{x}]$ is a valid continued fraction (since $x \neq 0$) and its value is $a_1 - x$, which, by Theorem 2.3, is $-1/M_2$. For $n \geq 2$, assume that the result for $n-1$, i.e.

$$m_n = \frac{-1}{[a_{n-1}, \ldots, a_1, -1/x]} \tag{2}$$

Now $[a_n,\ldots,a_1, -1/x]$ is a valid continued fraction since by assumption $[a_{n-1},\ldots,a_1, -1/x]$ is. By Theorem 2.3, $m_n = [a_n,m_{n+1}]$, therefore $-1/M_{n+1} = a_n - m_n = a_n + 1/[a_{n-1},\ldots,a_1,-1/x] = [a_n,\ldots,a_1, -1/x]$. //

We shall have occasion to use some results from elementary number theory. If p is an odd prime and $a \not\equiv 0 \pmod{p}$, then a is called a quadratic residue of p if there exists x such that $x^2 \equiv a \pmod{p}$.

1. -1 is a quadratic residue of p if and only if $p \equiv 1 \pmod{4}$.
   2 is a quadratic residue of p if and only if $p \equiv 1$ or $7 \pmod{8}$.
   -2 is a quadratic residue of p if and only if $p \equiv 1$ or $3 \pmod{8}$.
2. If N is a sum of two relatively prime squares, then N is a product of primes of the form $4k+1$ or twice such a product.

A proof of 1. may be found in [9], Chapter 5, Section 2. Result 2. is proved in [2], Chapter 20. (The converse of 2. is also true).

PART II

APPLICATIONS TO RATIONAL APPROXIMATIONS

Chapter 6. BEST APPROXIMATIONS TO REAL NUMBERS.

1. Introduction and Motivation.

This first section is meant to motivate and clarify the definitions of best approximation of the first and second kind.

Given a real number x, it is natural to look for a rational number which is a good approximation to x and yet whose denominator is "not too large". More precisely, (Problem A) given a real number x and a positive integer N, to find all pairs of integers a, b satisfying : (1) $1 \leq b \leq N$, and (2) for any integers r, s with $1 \leq s \leq N$, we have $|x-a/b| \leq |x-r/s|$. Clearly at least one such pair exists, and if one pair a, b is known, the others c, d (if any) can easily be found, since c/d must equal a/b or $2x - a/b$.

Definition. Rational number a/b $(b > 0)$ is said to be a best approximation of the first kind to x if $1 \leq d \leq b$, $c/d \neq a/b$ imply $| x - c/d | > | x - a/b |$.

The relation of this definition to Problem A is indicated by the following elementary propositions.

Proposition. Any best approximation c/d of the first kind is a solution of Problem A for some N, namely $N = d$.

Proof: Suppose $1 \leq s \leq d$. We must show that $|x-c/d| \leq |x-r/s|$. For $r/s = c/d$ this is immediate, while for $r/s \neq c/d$ it follows from the definition of best approximation of the first kind. //

Proposition: Except for the case $N = 1$ and $x = n + \frac{1}{2}$ (n an integer),

any solution a, b of Problem A with minimum b is a best approximation of the first kind to x.

Proof: Suppose $1 \leq d \leq b$, $c/d \neq a/b$, but $|x-c/d| \leq |x-a/b|$. Then $|x-c/d| = |x-a/b|$ (since a, b satisfies Problem A), $d = b$ (by minimality of b), and x is the average of a/b and $c/b = c/d$. Hence a and c differ by 1, for if t is strictly between a and C, t/b would be a better approximation to x.

Without loss of generality, assume $c = a + 1$. Write $a = qb+r$, $0<r<b$ ($r \neq 0$, for otherwise $a/b = q/1$, hence $b = 1$ by minimality of b, and then $x = a + 1/2$, contrary to assumption.) It is easily checked that

$$\frac{q\,(b-1) + r}{b-1}$$

exceeds a/b but not $(a+1)/b$, therefore is at least as good an approximation as a/b, contradicting the minimality of b. //

It may be noted that the situation is simplified if x is irrational : then clearly any solution of Problem A is a best approximation of the first kind.

Proposition: Exclude the case $N = 1$, $x = n + \frac{1}{2}$. Then any best approximation a/b of the first kind with $1 \leq b \leq N$ and such that b is maximum, is a solution of Problem A.

Proof: If a, b is not a solution of Problem A, let c, d be a solution with minimum d. We have $|x - c/d| < |x - a/b|$. Now by the preceding proposition, c/d is a best approximation of the first kind,

therefore by maximality of b, we have $d \leq b$. But then by definition of a/b being a best approximation, $|x-c/d| > |x - a/b|$, contradicting the first inequality. //

Thus (disregarding the trivial case $N = 1$, $x = n + \frac{1}{2}$), solution of Problem A for general N is equivalent to knowing all the best approximations of the first kind to x.

For some purposes it is desirable to consider the difference $|bx-a|$ rather than $|x - a/b|$. This leads to a second king of best approximation, as follows :

Definition. Rational number a/b $(b>0)$ is said to be a best approximation of the second kind to x if $1 \leq d \leq b$, $c/d \neq a/b$ imply $|dx-c| > |bx-a|$.

Theorem 6.1. Any best approximation of the second kind is also one of the first kind.

Proof: Multiplying $|dx-c| > |bx-a|$ by $1/d \geq 1/b$ gives $|x-c/d| > |x-a/b|$, as required. //

Examples are readily found to show that the converse of Theorem 6.1 is false. e.g. let $x = 5/12 - e$, where $0 < e \leq 1/60$; then $1/3$ is a best approximation of the first kind but not of the second, since it may be verified that $|2x - 1| \leq |3x - 1|$.

2. The Main Theorems.

We shall see that the convergents of the simple continued fraction expansion of x provide a very convenient means of finding all the best approximations of the first and second kind to x. We continue to use

the notation of Chapter 2 (formulas (1) to (7)): $x = [a_1,a_2,\ldots]$ is a finite or infinite simple continued fraction, with convergents $c_n = [a_1,a_2,\ldots,a_n] = p_n/q_n$. We begin by examining the difference between the continued fraction and its convergents.

Theorem 6.2. If x has at least $n + 2$ terms, then

$$| x - c_n | < \frac{1}{q_n q_{n+1}} \quad . \tag{1}$$

Proof: By the corollary to Theorem 2.1, x lies strictly between c_n and c_{n+1}, therefore $|x-c_n| < |c_n-c_{n+1}| = 1/q_n q_{n+1}$ (formula (6) of Chapter 2). //

It should be noted that (1) is replaced by an inequality if there are only $n + 1$ terms, for then $x = c_{n+1}$. Thus, if x has at least $n + 1$ terms, then $|x - c_n| \leq 1/q_n q_{n+1}$.

Theorem 6.3. Let $m_n = [a_n, a_{n+1},\ldots]$. Then :

$$x - c_n = \frac{(-1)^{n+1}}{q_n^2\left(m_{n+1} + \dfrac{q_{n-1}}{q_n}\right)} \tag{2}$$

Proof: $x = [a_1,\ldots,a_n, m_{n+1}]$

$$= \frac{p_n m_{n+1} + p_{n-1}}{q_n m_{n+1} + q_{n-1}} \quad .$$

Using this expression for x, we find that

$$x - \frac{p_n}{q_n} = \frac{p_{n-1}q_n - p_n q_{n-1}}{q_n^2 m_{n+1} + q_n q_{n-1}} ,$$

and (2) follows by use of formula (5) of Chapter 2. //

Theorem 6.4. If x has at least $n + 1$ terms and its last term (if there is one) is not 1, then :

$$| x - c_n | > \frac{1}{q_n (q_{n+1} + q_n)} . \tag{3}$$

Proof: Comparing (2) and (3), it is sufficient to prove that $q_n m_{n+1} + q_{n-1} < q_{n+1} + q_n$. Replacing q_{n+1} by $a_{n+1}q_n + q_{n-1}$, this becomes $m_{n+1} < a_{n+1} + 1$, which is clearly true. //

Theorem 6.5. Let x have at least $n + 2$ terms, with the last term (if any) not 1. Then :

$$|q_n x - p_n| > |q_{n+1} x - p_{n+1}| . \tag{4}$$

Proof: Since $a_{n+2} \geq 1$, we have $q_{n+1} + q_n \leq a_{n+2}q_{n+1} + q_n = q_{n+2}$. Therefore by Theorem 6.4, $|x - c_n| > 1/q_n q_{n+2}$, i.e.

$$|q_n x - p_n| > \frac{1}{q_{n+2}} . \tag{5}$$

But, as noted after Theorem 6.2, $|x - c_{n+1}| \leq 1/q_{n+1}q_{n+2}$, i.e.

$$\frac{1}{q_{n+2}} \geq |q_{n+1} x - p_{n+1}| . \tag{6}$$

(4) follows from (5) and (6). //

Corollary: Except for the convergents c_1, c_2 of $[\, a_1, 1, 1 \,]$, any convergent c_{n+1} is closer to x than the preceding convergent c_n, i.e.

$$| x - c_n | > | x - c_{n+1} | . \qquad (7)$$

Proof: (i) If there are only $n + 1$ terms then

$$| x - c_{n+1} | = 0 < 1/q_n q_{n+1} = | c_{n+1} - c_n | = | x - c_n | .$$

(ii) If there are only $n + 2$ terms, then $x = c_{n+2}$, and using formula (7) of Chapter 2,

$$| x - c_n | = \frac{a_{n+2}}{q_n \, q_{n+2}} .$$

also,

$$| x - c_{n+1} | = \frac{1}{q_{n+1} \, q_{n+2}}$$

therefore (7) is equivalent to $a_{n+2} \, q_{n+1} > q_n$. Now $q_{n+1} \geq q_n$ (equality only if $n = 1$ and $a_2 = 1$), hence (7) is false only for $n = 1$, $a_2 = 1$, $a_3 = 1$; since the present case is assuming only $n+2$ terms, this means that $x = [\, a_1, 1, 1 \,]$, i.e. the exception noted.

(iii) If there are $n + 3$ terms or more, we can absorb any final unit into the second to last term, therefore we always have (4), by Theorem 6.5. Multiplying by $1/q_n \geq 1/q_{n+1}$ gives (7). //

Assuming b and d positive, the mediant (also called the median value) of a/b and c/d is defined to be (a+c) / (b+d), and is, as one can trivially verify, strictly between a/b and c/d if

the latter are distinct. Consider the fractions

$$\frac{p_n + ip_{n+1}}{q_n + iq_{n+1}} \qquad (0 \leq i \leq a_{n+2}) . \qquad (8)$$

We observe that the first of these is c_n, the last is c_{n+2}, and each after the first is the mediant of the preceding one and p_{n+1}/q_{n+1}. The fractions (8), other than the first and last, are called intermediate fractions (or secondary convergents) of the continued fraction. The reason for introducing intermediate fractions is the following result : every best approximation of the first kind to x is either a convergent or an intermediate fraction of the simple continued fraction expansion of x. Since the proof is somewhat long, but not difficult, and the result will not be needed in what follows, we shall omit the proof ; it may be found in [4], § 6, Theorem 15. As will be shown, every convergent is a best approximation of the first kind ; however, this is not true of every intermediate fraction. For best approximations of the second kind, a much sharper result (presented in the next two theorems) is true, and this constitutes the main reason for considering best approximations of the second kind.

Theorem 6.6. If a/b is a best approximation of the second kind, then there exists $n \geq 1$ such that $a/b = c_n$.

Proof: Let $L \leq \infty$ be the number of terms in the given continued fraction. For the purposes of this proof, " c, d gives a contradiction " will mean that $1 \leq d \leq b$, $c/d \neq a/b$, and $| dx-c | \leq | bx-a |$, thus contradicting that a/b is a best approximation of the second kind. We shall have

occasion to use the fact that for any integers e, f, g, h (f,h positive), $e/f \neq g/h$ implies $|e/f - g/h| \geq 1/fh$; this follows since $|eh-fg|$ is not zero, hence at least 1.

If $L = 1$, then $x = a_1 = c_1$, and $a/b = c_1$, for otherwise a_1, 1 gives a contradiction. Therefore assume $L \geq 2$. If $a/b < c_1$, we have $|x-a_1| < |x-a/b| \leq b\,|x-a/b| = |bx-a|$, so that a_1, 1 gives a contradiction.

If $a/b > c_2$, then

$$| x - \frac{a}{b} | \geq | c_2 - \frac{a}{b} | \geq \frac{1}{b\, q_2}$$

i.e. $$| b_x - a | \geq \frac{1}{q_2} \quad ;$$

but $| x - a_1 | \leq |c_2 - a_1 | = 1/q_2$, therefore a_1, 1 again gives a contradiction. Thus, assuming a/b is not equal to any convergent, we have $c_1 < a/b < c_2$; also, $a/b \neq x$, for otherwise x is rational and $a/b = c_L$. Therefore we are clearly in one of two cases :

(i) For some $n \geq 1$, a/b is strictly between c_n and c_{n+2}.

(ii) L is finite, and a/b is strictly between c_{L-1} and $c_L = x$.

We note that if a/b is strictly between c_r and c_{r+1} , then

$$\frac{1}{bq_r} \leq | \frac{a}{b} - \frac{p_r}{q_r} | < | c_r - c_{r+1} | = \frac{1}{q_r q_{r+1}} \quad ,$$

therefore $q_{r+1} < b$.

In case (ii), therefore, $q_L < b$; also, $|q_L x - p_L| = 0$, hence

p_L, q_L gives a contradiction.

In case (i), we show that p_{n+1}, q_{n+1} gives a contradiction. Now a/b is also strictly between c_n and c_{n+1}, hence $q_{n+1} < b$. Furthermore, a/b is clearly at least as close to c_{n+2} as it is to x, so that

$$| x - \frac{a}{b} | \geq | \frac{p_{n+2}}{q_{n+2}} - \frac{a}{b} | \geq \frac{1}{b\, q_{n+2}}$$

i.e. $$| bx - a | \geq \frac{1}{q_{n+2}} \quad .$$

But $|q_{n+1} x - p_{n+1}| = q_{n+1} | x - c_{n+1} | \leq 1/q_{n+2}$ (from Theorem 6.2). Therefore $|q_{n+1} x - p_{n+1}| \leq |b_{x-a}|$. //

Theorem 6.7 Assume that the continued fraction used for x does not have 1 as a last term. Let $n \geq 1$, and exclude the following cases :

1. $n = 1$, there are at least 2 terms, and $a_2 = 1$.
2. $n = 1$, and the fraction is $[a_1, 2]$.

Then the convergent p_n/q_n is a best approximation of the second kind (and hence also of the first kind).

Remarks: (i) Suppose $n \geq 2$, and the last term is $a_{n+1} = 1$; then p_n/q_n is not a best approximation of the second kind ; $c_{n-1} \neq c_n$, $1 \leq q_{n-1} \leq q_n$, and (since $x = c_{n+1}$) , $|q_{n-1} x - p_{n-1}| = |q_n x - p_n|$, each being $1/q_{n+1}$.

(ii) In case 1, or 2, above, we have $a_1 + \frac{1}{2} \leq x \leq a_1 + 1$, therefore $c_1 = a_1/1$ is not a best approximation of the first or second

kind (by comparison with $(a_1 + 1) / 1$) .

Proof of Theorem 6.7: Let $m \leq \infty$ be the number of terms in the continued fraction. First of all, we can assume $n < m$, for if $n = m$, then $p_n/q_n = x$ and the result is immediate.

Let X be the set of all pairs (u, v) where u may be any integer and $v = 1, 2, \ldots, q_n$. Define a function F with domain X by $F(u,v) = | vx - u |$. Letting t be any value of F, $F(u,v) \leq t$ is equivalent to $vx - t \leq u \leq vx + t$; v is bounded, therefore only a finite (and nonzero) number of elements (u, v) satisfy $F(u, v) \leq t$; hence F attain a minimum at one of these elements. Let M be the set of all elements of X at which this minimum is attained, and let (u_o, v_o) be an element of M with least v_o.

Then we have :

I. For all (u, v) in X, $F(u,v) \geq F(u_o, v_o)$.

II. If $F(u,v) \leq F(u_o,v_o)$, then (u, v) is in M, and $v \geq v_o$.

We also claim the following :

III. If (u, v_o) is in M, then $u = u_o$.

Once III is proved, it will follow that u_o/v_o is a best approximation of the second kind to x, for if $1 \leq d \leq v_o$ and $F(c,d) \leq F(u_o,v_o)$, then II gives $d = v_o$ and III gives $c = u_o$.

To prove III, suppose (u, v_o) is in M, but $u \neq u_o$. Then $| v_o x - u | = | v_o x - u_o |$ and

$$x = \frac{u + u_o}{2 v_o} .$$

also, $|v_o x - u_o| \neq 0$, hence $| v_o x - u_o| = | (u + u_o) / 2 - u_o/1 | \geq \frac{1}{2}$;

x is rational, so m is finite and $x = c_m$. Now $u + u_o$ and $2v_o$ are relatively prime, for suppose $u+u_o = kp$, $2v_o = kq$, $k \geq 2$; then $q \leq v_o$, so that (p, q) is in X, and $|qx - p| = 0 < |v_o x - u_o|$ contradicts I. Therefore, $u + u_o - p_m$, $2v_o = q_m$; $q_m > 1$ shows that $m \geq 2$. Now $|q_{m-1} x - p_{m-1}| = q_{m-1}|c_m - c_{m-1}|$ $= 1/q_m = 1/2v_o \leq 1/2 \leq |v_o x - u_o|$, therefore II gives a contradiction if we can show that $q_{m-1} < v_o$. Now $2v_o = a_m q_{m-1} + q_{m-2}$ and we assumed $a_m \geq 2$, therefore $v_o > q_{m-1}$ unless $a_m = 2$ and $m = 2$ (hence also $n = 1$, since we dismissed $n = m$); this case was excluded, therefore III is proved.

Theorem 6.6 now shows that there exists $s \geq 1$ such that $u_o/v_o = c_s$. If $u_o = kp$ and $v_o = kq$, then $|qx-p| \leq |v_o x - u_o|$, and II gives $q \geq v_o$, so that $k = 1$ and u_o, v_o are relatively prime. Thus, $u_o = p_s$, $v_o = q_s$. We complete the proof by showing that $s = n$. Since we excluded the case $n = 1$ and $a_2 = 1$, $n < s$ leads to $q_n < q_s$, which is false by definition of X. Suppose $s < n$; then $s + 1 \leq n$, hence $q_s + q_{s+1} \leq q_{n-1} + q_n$; using I, Theorems 6.2 and 6.4, and recalling $n < m$:

$$\frac{1}{q_{n+1}} \geq |q_n x - p_n| \geq |q_s x - p_s|$$

$$> \frac{1}{q_s + q_{s+1}} \geq \frac{1}{q_{n-1} + q_n}$$

Therefore $q_{n+1} < q_n + q_{n-1}$, which is false since $q_{n+1} = a_{n+1} q_n + q_{n-1}$. //

3. A Restatement.

The results proved in this section are closely related to Theorems 6.6 and 6.7, but are in a form which is sometimes more convenient to use. As demonstrated by Niven and Zuckerman [6] in their Theorem 7.13, they may also be proved directly without using our Theorems 6.6 and 6.7.

Theorem 6.8. Let $n \geq 1$ and assume that the fraction used for x has at least $n + 1$ terms. Then $b > 0$, $|bx-a| < |q_n x-p_n|$ imply $b \geq q_{n+1}$.

Proof: We can assume a and b relatively prime, for if $a = k_c$, $b = k_d$, $(c, d) = 1$, we could apply the theorem to c, d to get $d \geq q_{n+1}$, therefore $b \geq q_{n+1}$.

If $n = 1$ and $a_2 = 1$, the result is trivial, since then $q_{n+1}=1$. If $n = 1$ and the fraction is $[a_1, 2]$, we have $b \geq 2 = q_{n+1}$ for otherwise $b = 1$ and $|bx-a| \geq \frac{1}{2} = |q_1 x - p_1|$, contradicting the assumption. Also, we can assume that the continued fraction does not have 1 as a last term, for suppose the number of terms is m and $a_m = 1$. If $m \geq n + 3$ we can absorb a_m into a_{m-1} and the theorem still gives $b \geq q_{n+1}$. If $m = n + 2$, absorbing a_{n+2} into a_{n+1}, we get (where primes refer to the new fraction) $b \geq q'_{n+1} = (a_{n+1}+1)\, q_n + q_{n-1} = q_{n+1} + q_n > q_{n+1}$. Finally, if $m = n + 1$, we can assume $n > 1$ (since the case $n = 1$, $a_2 = 1$ has been treated); as indicated in Remark (i) following the statement of Theorem 6.7, $|q_{n-1}x-p_{n-1}| = |q_n x - p_n|$; therefore absorbing a_{n+1} into a_n, we have $b \geq q'_n$; but $q'_n = (a_n+1)\, q_{n-1} + q_{n-2} = q_n + q_{n-1} = q_{n+1}$.

Thus, Theorem 6.7 gives that p_n/q_n is a best approximation of the second kind. Also, $a/b \neq p_n/q_n$ (otherwise $|b_x-a| = |q_nx-p_n|$, since $(a, b) = 1)$, therefore $b \leq q_n$ implies $|b_x-a| > |q_nx-p_n|$, which is false. This shows $b > q_n$. Suppose $b < q_{n+1}$. Then a/b is not a convergent, hence by Theorem 6.6 it is not a best approximation of the second kind. Thus there exist c, d such that $1 \leq d \leq b$, $a/b \neq c/d$, $(c, d) = 1$, and $|dx-c| \leq |bx-a|$. We claim that $d \neq b$, for $d = b$ gives $a \neq c$ and $1 \leq |a-c| \leq |dx-c| + |bx-a| \leq 2|bx-a|$, hence $|bx-a| \geq \frac{1}{2}$; but except in the case $n = 1$, $a_2 = 1$, which has been treated) $|q_nx - p_n| \leq 1/q_{n+1} \leq \frac{1}{2}$, and $|bx-a| < |q_nx-p_n|$ is contradicted. Therefore $d < b$. Now if $d > q_n$, then c/d is not a best approximation of the second kind and $|dx-c| < |q_nx-p_n|$, so that by the same argument, we get e/f, $1 \leq f < d$, $(e,f) = 1$, $|fx-e| \leq |dx-c|$. This process may be continued until we eventually obtain r/s, $1 \leq s \leq q_n$, $(r,s) = 1$, and $|sx-r| < |q_nx-p_n|$ (hence $r/s \neq p_n/q_n$), contradicting that p_n/q_n is a best approximation of the second kind. Therefore $b \geq q_{n+1}$. //

Corollary: Let $n \geq 1$ and assume that the fraction used for x has at least n terms. Exclude the case $n = 1$, $a_2 = 1$. Then $b > o$, $|x-a/b| < |x-p_n/q_n|$ imply $b > q_n$.

Proof: Assume at least $n + 1$ terms, since the result is empty for n terms. If $b \leq q_n$, multiplying the given inequality by this gives $|bx-a| < |q_nx-p_n|$, therefore $b \geq q_{n+1}$, by the theorem. This is a contradiction, since $q_{n+1} > q_n$. //

Chapter 7. HURWITZ'S THEOREM AND RELATED RESULTS.

Since $q_{n+1} \geq q_n$, Theorem 6.2 shows that $|x - c_n| < 1/q_n^2$, and therefore for any irrational number x there exist infinitely many rationals a/b such that $|x-a/b| < 1/b^2$. Let us now consider the possibility of replacing b^2 by some other function of b. More specifically, for how large a value of k can b^2 be replaced by kb^2 ? Investigation of this problem leads to the striking result, first proved by Hurwitz in 1891 ([3] , using Farey sequences rather than continued fractions), that k can be as large as $\sqrt{5}$, but no larger. First, however, we shall examine the easier case $k = 2$.

Theorem 7.1. Exclude the case where $n = 1$ and the fraction is $[a_1,1,1]$. Given any two consecutive convergents c_n and c_{n+1}, at least one of the following two inequalities is true :

$$| x - c_n | < \frac{1}{2q_n^2} \qquad (1)$$

$$|x - c_{n+1}| < \frac{1}{2q_{n+1}^2} \qquad (2)$$

Proof: x lies between c_n and c_{n+1}, therefore $|x-c_{n+1}| = |c_{n+1}-c_n| - |x - c_n| = 1/q_nq_{n+1} - |x - c_n|$. Assuming (1) false, we obtain

$$| x - c_{n+1}| \geq \frac{1}{q_nq_{n+1}} - \frac{1}{2q_n^2} . \qquad (3)$$

Now $1/ab - 1/2a^2 < 1/2b^2$ is equivalent to $2ab-b^2 < a^2$, i.e. $(a-b)^2 > 0$, i.e. $a \neq b$. But $q_{n+1} \neq q_n$, proving (2), unless $n = 1$ and $a_2 = 1$.

If $n = 1$ and $a_2 = 1$, then $c_n = a_1$, $c_{n+1} = a_1 + 1$, $q_n = 1 = q_{n+1}$, and clearly one of (1), (2) is true unless $x = a_1 + \frac{1}{2}$, in which case the fraction must be $[a_1, 1, 1]$. Since this case was excluded, the theorem is proved. //

Corollary: Given any irrational number x, there are infinitely many rational numbers a/b such that

$$| x - a/b | < \frac{1}{2b^2} . \tag{4}$$

Theorem 7.2. If (4) holds, then a/b is a convergent.

Proof: We can assume $b > o$. Given (4), we prove that a/b is a best approximation of the second kind to x; the desired result then follows by Theorem 6.6.

Suppose $1 \leq d \leq b$ and $c/d \neq a/b$; we must prove

$$|dx-c| > |bx-a| . \tag{5}$$

By (4), it is enough to show $| dx-c | \geq 1/2b$.

$$\begin{aligned} \frac{1}{db} \leq \left|\frac{c}{d} - \frac{a}{b}\right| &\leq \left| x - \frac{c}{d} \right| + \left| x - \frac{a}{b} \right| \\ &< \left| x - \frac{c}{d} \right| + \frac{1}{2b^2} \\ &\leq \left| x - \frac{c}{d} \right| + \frac{1}{2bd} . \end{aligned}$$

Therefore $\left| x - \frac{c}{d} \right| > \frac{1}{db} - \frac{1}{2db} = \frac{1}{2db}$,

i.e. $|dx - c| > 1/2b$. //

Theorem 7.3. If $k > \sqrt{5}$, then there is an irrational number x

(for example $x = (1 + \sqrt{5})/2$) such that

$$\left| x - \frac{a}{b} \right| < \frac{1}{kb^2} \tag{6}$$

is true for only a finite number of rational numbers a/b.

Proof: Let $x = (1 + \sqrt{5}) / 2$ and let $F_1, F_2, F_3, \ldots$ be the Fibonacci numbers $1,1,2,\ldots$ As was shown in Chapter 4, $x = [1, 1, 1, \ldots]$ and $p_n = F_{n+1}$, $q_n = F_n$. By Theorem 6.3,

$$\left| x - c_n \right| = \frac{1}{q_n^2 \left(x + \frac{F_{n-1}}{F_n}\right)} , \tag{7}$$

But $x + F_{n-1}/F_n \to x + 1/x = \sqrt{5}$. Therefore $k > \sqrt{5}$ implies that for all n sufficiently large, $|x - c_n| > 1/kq_n^2$. But $k > 2$, hence by Theorem 7.2, if $|x - a/b| < 1/kb^2$ then a/b is a convergent. Therefore the theorem is proved. //

In preparation for proving the next theorem, we introduce some notation :

$$m_n = [a_n, a_{n+1}, \ldots] \qquad (n \geq 1) \tag{8}$$

$$u_n = \frac{q_{n-2}}{q_{n-1}} \qquad (n \geq 2) \tag{9}$$

$$w_n = u_n + m_n \qquad (n \geq 2) \tag{10}$$

We observe that $m_n = a_n + 1/m_{n+1}$, and

$$\frac{1}{u_{n+1}} = \frac{q_n}{q_{n-1}} = \frac{a_n q_{n-1} + q_{n-2}}{q_{n-1}} = a_n + u_n .$$

Therefore $$\frac{1}{u_{n+1}} + \frac{1}{m_{n+1}} = w_n \quad (n \geq 2) \tag{11}$$

and $$a_n = \frac{1}{u_{n+1}} - u_n \qquad (n \geq 2) \; . \tag{12}$$

The formula proved in Theorem 6.3 becomes

$$| x - c_n | = \frac{1}{q_n^2 \, w_{n+1}} \; . \tag{13}$$

Lemma: Let $n \geq 2$. From $w_n \leq \sqrt{5}$ and $w_{n+1} \leq \sqrt{5}$ it follows that $u_{n+1} > (\sqrt{5} - 1) / 2$.

Proof: Using (11) , we have

$$\frac{1}{u_{n+1}} + \frac{1}{m_{n+1}} \leq \sqrt{5} \; , \quad u_{n+1} + m_{n+1} \leq \sqrt{5} \; .$$

Therefore $(\sqrt{5} - u_{n+1}) (\sqrt{5} - \frac{1}{u_{n+1}}) \geq m_{n+1} \frac{1}{m_{n+1}} = 1$.

Multiplying by u_{n+1} and completing the square,

$$(u_{n+1} - \frac{\sqrt{5}}{2})^2 \leq \frac{1}{4}$$

$$| u_{n+1} - \frac{\sqrt{5}}{2} | \leq \frac{1}{2}$$

$$u_{n+1} \geq \frac{\sqrt{5}}{2} - \frac{1}{2} = \frac{\sqrt{5}-1}{2} \; .$$

But u_{n+1} is rational, therefore the lemma is proved. //

Theorem 7.4. Let $n \geq 1$ and assume that $x = [a_1, a_2, \ldots]$ has at least $n + 2$ terms. Then the inequality

$$| x - c_i | < \frac{1}{\sqrt{5} \, q_i^2} \tag{14}$$

is true for at least one of the three values $i = n$, $i = n + 1$, $i = n + 2$.

Proof: We can assume at least $n + 3$ terms, for otherwise (14) is true

for $i = n + 2$. Assume (14) false for all three values of i. Then by (13), $w_i \leq \sqrt{5}$ $(i = n+1, n+2, n+3)$, therefore by the lemma, $u_i > (\sqrt{5} - 1)/2$ $(i = n+2, n+3)$. But then (12) gives

$$a_{n+2} < \frac{2}{\sqrt{5}-1} - \frac{\sqrt{5}-1}{2} = 1 ,$$

which is false. //

Corollary: Given any irrational number x, there are infinitely many rational numbers a/b such that

$$\left| x - \frac{a}{b} \right| < \frac{1}{\sqrt{5}b^2} .$$

PART III

APPLICATIONS TO NUMBER THEORY

Chapter 8. THE NUMBER OF STEPS IN THE EUCLIDEAN ALGORITHM

In this chapter we prove an elementary yet significant result concerning the Euclidean algorithm. The ease of the proof is a good illustration of the power of the theory of continued fractions.

Let r, s be any integers, with $s > 0$. Consider the Euclidean algorithm for r and s :

$$1. \quad r = a_1 s + r_1 \qquad (0 < r_1 < s)$$

$$2. \quad s = a_2 r_1 + r_2 \qquad (0 < r_2 < r_1)$$

$$3. \quad r_1 = a_3 r_2 + r_3 \qquad (0 < r_3 < r_2)$$

$$\vdots$$

$$n-1. \quad r_{n-3} = a_{n-1} r_{n-2} + r_{n-1} \qquad (0 < r_{n-1} < r_{n-2})$$

$$n. \quad r_{n-2} = a_n r_{n-1} .$$

Here, $n \geq 1$ is called the number of steps, and we shall use the notation $E(r,s) = n$.

As shown in Theorem 3.3, $[a_1, a_2, \ldots, a_n]$ is the simple continued fraction expansion of r/s, and (if $n > 1$) $a_n \geq 2$. Therefore we can write :

$$\frac{r}{s} = [a_1, a_2, \ldots, a_{n-1}, a_n - 1, 1] \qquad (1)$$

Using F_i for the Fibonacci numbers and letting p_i and q_i refer to (1), we have (Theorem 4.1) $q_{n+1} \geq F_{n+1}$. But $r/s = p_{n+1}/q_{n+1}$, therefore $s \geq F_{n+1}$. Now suppose $s < F_m$. Then $F_{n+1} \leq s < F_m$, therefore $n + 1 < m$, i.e. $E(r,s) \leq m - 2$. This proves the following

theorem :

Theorem 8.1. Let $F_1 = 1$, $F_2 = 1,\ldots$ be the Fibonacci numbers. If r is any integer and $0 < s < F_m$, then the number of steps in the Euclidean algorithm for r and s is at most $m-2$.

Example 1. Taking $a = F_{m-1}$ and $r = F_m$, we have $F_m/F_{m-1} = [1,1,\ldots,1]$ ($m-1$ 1's), hence $E(F_m, F_{m-1}) = m-2$.

Example 2. Let $r = 18$, $s = 11$. 11 is between the Fibonacci numbers $F_6 = 8$ and $F_7 = 13$, hence the theorem predicts $E(18,11) \leq 7-2 = 5$. In fact, $E(18,11) = 5$:

$$18 = 1.11 + 7$$

$$11 = 1.7 + 4$$

$$7 = 1.4 + 3$$

$$4 = 1.3 + 1$$

$$3 = 3.1$$

Chapter 9. PERIODIC SIMPLE CONTINUED FRACTIONS

Periodic simple continued fractions have many interesting and useful properties, due primarily to the fact that a continued fraction is periodic if and only if it represents a quadratic irrational. This striking result was first proved by Lagrange in 1770. In particular, the simple-continued-fraction expansion of $\sqrt{D}$ where D is a positive nonsquare integer) provides the key to the solution of Pell's equation $x^2 - Dy^2 = \pm 1$, to be discussed in later chapters.

By a quadratic <u>irrational</u> we mean a number of the form $A+B\sqrt{D}$, where A and B are rational numbers $(B \neq 0)$ and D is a positive nonsquare integer. We note that if $A + B\sqrt{D} = E + F\sqrt{D}$, then $A = E$ and $B = F$ (for if $B \neq F$, we would have $\sqrt{D} = (E-A) / (B-F)$, contradicting that $\sqrt{D}$ is irrational ; hence $B = F$ and consequently $A = E$). The <u>conjugate</u> of $x = A + B\sqrt{D}$ is defined to be $\bar{x} = A - B\sqrt{D}$. One easily verifies that, if $y = E + F\sqrt{D}$, then $\overline{x + y} = \bar{x} + \bar{y}$ and $\overline{xy} = \bar{x}\,\bar{y}$ hence also $\overline{x^{-1}} = \bar{x}^{-1}$) ; in other words, the operation of taking the conjugate is an automorphism of the field $Q(\sqrt{D})$ (i.e. the field of elements $A + B\sqrt{D}$, where A and B are rational).

Clearly a root of a quadratic equation with integer coefficients and positive nonsquare discriminant is a quadratic irrational.

Furthermore, it is not difficult to see that any given quadratic irrational is a root of precisely one quadratic equation $ax^2+bx+c = 0$ where a,b,c are integers, $a > 0$, and $(a,b,c) = 1$. To prove the last statement , $x = A + B\sqrt{D}$ is a root of $x^2-2Ax + (A^2-B^2D) = 0$, which

can clearly be put into the required form ; to show uniqueness, suppose $ax^2 + bx + c = 0 = dx^2 + ex + f$ ($a, d > 0$, $(a,b,c) = 1 = (d,e,f)$) ; eliminating x^2 , we get $(bd-ae)\,x = -(cd-af)$, hence (since x is irrational) $bd = ae$ and $cd = af$; assuming first that $e \neq 0$ and $f \neq 0$, we have $\frac{a}{d} = \frac{b}{e} = \frac{c}{f}$; denoting each fraction by k and letting $rd+se+tf = 1$, we get $k = ra+sb+tc$, showing that k is an integer ; but $(a,b,c) = 1$, therefore in fact $k = 1$, hence $a = d$, $b = e$, $c = f$; the case $e = 0$ or $f = 0$ is treated similarly.

One half of Lagrange's result is relatively easy, and we state it in the following theorem :

Theorem 9.1 The value of any periodic simple continued fraction is a quadratic irrational.

Proof: Let $x = [a_1, a_2, \ldots]$, where, for all $n > m$, $a_{n+r} = a_n$ (here, $m \geq 0$ and $r \geq 1$). Let $y = [a_{m+1}, a_{m+2}, \ldots]$. Then $x = [a_1, \ldots, a_m, y] = [a_1, \ldots, a_{m+r}, y]$, therefore

$$x = \frac{p_m y + p_{m-1}}{q_m y + q_{m-1}} = \frac{p_{m+r} y + p_{m+r-1}}{q_{m+r} y + q_{m+r-1}} .$$

Hence y satisfies a quadratic equation with integer coefficients (the leading coefficient $p_m q_{m+r} - q_m p_{m+r}$ is not zero, since $p_m/q_m \neq p_{m+r}/q_{m+r}$ if $m > 0$, while $p_m q_{m+r} = 1 \cdot q_{m+r} \neq 0 = q_m p_{m+r}$ if $m = 0$). Therefore y, and consequently x, is a quadratic irrational. //

As a first step towards proving the converse, we shall determine which real numbers have purely periodic expansions. These turn out to be the reduced quadratic irrational, defined as follows :

Definition: A reduced quadratic irrational is a quadratic irrational x such that $x > 1$ and $-1 < \bar{x} < 0$.

Theorem 9.2 Any purely periodic simple continued fraction

$$x = [\overline{a_1, a_2, \ldots, a_r}] \tag{1}$$

is a reduced quadratic irrational, and $\bar{x}y = -1$, where

$$y = [\overline{a_r, a_{r-1}, \ldots, a_1}] . \tag{2}$$

Proof: $x > a_1 \geq 1$, and by Theorem 9.1 x is a quadratic irrational. Applying Theorem 5.2, we have

$$-\frac{1}{x} = [a_r, a_{r-1}, \ldots, a_1, -\frac{1}{x}] . \tag{3}$$

(2) and (3) show that when the proof of Theorem 9.1 is applied to y and $-\frac{1}{x}$ (using $m = 0$) , these numbers satisfy the same quadratic equation. But y and $-1/x$ are distinct (they are of opposite sign), therefore each is the conjugate of the other, proving that $\bar{x}y = -1$. Finally, $\bar{x} = -1/y$ and $y > 1$, therefore $-1 < \bar{x} < 0$. //

Lemma 1. Any quadratic irrational x may be expressed in the form

$$x = \frac{P + \sqrt{D}}{Q} , \tag{4}$$

where P, Q, D are integers, $D > 0$, and D is not a square. For any such representation, x is reduced if and only if

$$P < \sqrt{D} \tag{5}$$

$$Q-P < \sqrt{D} \tag{6}$$

$$P+Q > \sqrt{D} \tag{7}$$

Given positive nonsquare D, there are precisely h(h+1) reduced quadratic irrational of the form (4), where $h = [\sqrt{D}]$ (integral part of $\sqrt{D}$).

Proof: $\frac{a}{b} + \frac{c}{d}\sqrt{k}$ can be written as $(ad + \sqrt{b^2c^2k}) / (bd)$.

Assume (4). (i) let x be reduced, i.e.

$$\frac{P + \sqrt{D}}{Q} > 1 \tag{8}$$

$$\frac{P - \sqrt{D}}{Q} < 0 \tag{9}$$

$$\frac{P - \sqrt{D}}{Q} > -1 . \tag{10}$$

Suppose $Q < 0$; then (8) shows $P < 0$; but then $P - \sqrt{D} < 0$, contradicting (9). Therefore $Q > 0$, and (5)-(7) follow easily from (8)-(10). (ii) Let (5)-(7) be satisfied.
Subtracting (5) and (7) gives $Q > 0$, so that (8) - (10) follow.

Geometrically, (5) - (7) mean that (P,Q) is a lattice point strictly inside the triangle whose vertices are $(\sqrt{D}, 0)$, $(0,\sqrt{D})$, $(\sqrt{D}, 2\sqrt{D})$. Clearly the number of such lattice points is $2(1+2+\ldots+h) = h(h + 1)$. //

Lemma 2. If x is a reduced quadratic irrational, then so is $-1/\bar{x}$.

Proof: $-1 < \bar{x} < 0$, therefore $-1/\bar{x} > 1$; $x > 1$, hence $-1/x$, which is the conjugate of $-1/\bar{x}$, lies between -1 and 0. //

Lemma 3. Let x be a quadratic irrational satisfying

$$ax^2 + bx + c = 0 \qquad (11)$$

where a ($\neq 0$), b, c are integers, and let k and D be any integers such that $b^2-4ac = k^2D$ and k divides $2a$, b, and $2c$. Then given any sequence of integers $s_1, s_2, \ldots$, the numbers $x_1 = x$ and

$$x_{i+1} = \frac{1}{x_i - s_i} \qquad (i \geq 1) \qquad (12)$$

are of the form

$$x_i = \frac{P_i + \sqrt{D}}{Q_i} \qquad (i \geq 1) \qquad (13)$$

where P_i and Q_i are integers.

Proof: Given (11) and the assumption about a,b,c, it is clear that $x = (-b \pm \sqrt{k^2D}) / 2a$ is of the form (13). Thus, it is sufficient to prove that each x_i satisfies an equation $a_i x_i^2 + b_i x_i + c_i = 0$ ($a_i \neq 0$) where $b_i^2 - 4a_ic_i = k^2D$ and k divides $2a_i$, b_i, c_i. For this, it is enough to prove that $y = 1/(x-s)$ (s any integer) does. Now $x = s + 1/y$, and substituting this into (11) and simplifying, we find $dy^2 + ey + f = 0$, where $d = as^2 + bs + c$, $e = 2as + b$, and $f = a$, also, one can check that $e^2 - 4df = b^2 - 4ac = k^2D$; clearly k divides $2d$, e, $2f$; finally, $d \neq 0$, for otherwise $e^2 = k^2D$, contradicting that x is irrational. //

Theorem 9.3. The simple-continued-fraction expansion of any reduced quadratic irrational is purely periodic. Furthermore, if the quadratic irrational is $x = [a_1, a_2, \ldots]$ and satisfies $ax^2 + bx + c = 0$ where $a(\neq 0)$, b, c are integers with $b^2 - 4ac = k^2D$ and k dividing

2a, b, and 2c, then each $m_n = [a_n, a_{n+1}, \ldots]$ $(n \geq 1)$ is a reduced quadratic irrational of the form $(P_n + \sqrt{D}) / Q$. (P_n, Q_n integers), and the fundamental period does not exceed $h(h+1)$, where $h = [\sqrt{D}]$.

Proof: In lemma 3, take $s_i = a_i$. Then $x_i = m_i$ and therefore each m_n is of the form $(P_n+\sqrt{D}) / Q_n$. Assume m_n is reduced. Now $m_{n+1} > a_{n+1} \geq 1$; also

$$\bar{m}_{n+1} = \frac{1}{\bar{m}_n - a_n} \quad ; \tag{14}$$

but $0 < -\bar{m}_n < 1$, i.e. $a_n < a_n - \bar{m}_n < a_n + 1$, hence by (14) $1/(a_n+1) < -\bar{m}_{n+1} < 1/a_n$; now $a_n \geq 1$ (for $n = 1$ this follows from the fact that $x > 1$, since x is reduced), therefore m_{n+1} is reduced. But $x = m_1$ is reduced, therefore every m_n is reduced. By lemma 1, there are only $h(h+1)$ reduced quadratic irrational of the form (4), hence there exist $r \geq 1$ and $t \geq 1$ such that $m_{r+t} = m_r$ and $t \leq h(h+1)$. Choose the smallest possible r. Then $r = 1$, for suppose $r > 1$. $m_{r-1} = a_{r-1} + 1/m_r$, so that, denoting $-1/\bar{m}_n$ by b_n, we have $b_r = a_{r-1} + 1/b_{r-1}$. Similarly $b_{r+t} = a_{r+t-1} + 1/b_{r+t-1}$. By lemma 2, b_n is reduced, hence $b_n > 1$, therefore $[b_r] = a_{r-1}$, $[b_{r+t}] = a_{r+t-1}$. Since $b_r = b_{r+t}$, this shows that $a_{r-1} = a_{r+t-1}$, and consequently $m_{r-1} = m_{r+t-1}$, contradicting the minimality of r. Thus $m_1 = m_{1+t}$, proving that x is purely periodic with period $t \leq h(h+1)$. //

The following result completes the proof of Lagrange's theorem.

Theorem 9.4. The simple-continued-fraction expansion of any quadratic irrational is periodic. In fact, if the quadratic irrational is $x = [a_1, a_2, \ldots]$ and satisfies $ax^2+bx+c = 0$ where a, b, c are integers with $b^2-4ac = k^2D$ and k dividing $2a$, b, and $2c$, then each $m_n = [a_n, a_{n+1}, \ldots]$ is of the form $[P_n + \sqrt{D})/Q_n$ (P_n, Q_n integers), and all m_n from some point onward are reduced. The fundamental period does not exceed $h(h+1)$, where $h = [\sqrt{D}]$.

Proof: As in the proof of Theorem 9.3, the fact that $m_n = (P_n+\sqrt{D})/Q_n$ follows from lemma 3. In view of Theorems 9.3 and 9.2, it remains only to find an n for which m_n is reduced. (The fact that the period does not exceed $h(h+1)$ is clear from lemma 1.). Since $x = [a_1,\ldots,a_n,m_{n+1}]$, we have

$$x = \frac{m_{n+1}\, p_n + p_{n-1}}{m_{n+1}\, q_n + q_{n-1}} \qquad (n \geq 0) \qquad (15)$$

Taking conjugates and solving for $\bar{m}_{n+1}$,

$$\bar{m}_{n+1} = -\frac{\bar{x}\, q_{n-1} - p_{n-1}}{\bar{x}\, q_n - p_n} = -\frac{q_{n-1}}{q_n} \cdot \frac{\bar{x} - c_{n-1}}{\bar{x} - c_n} \qquad (16)$$

Now as n increases, each convergent is alternately smaller and larger than the previous one, therefore $(\bar{x} - c_{n-1}) / (\bar{x} - c_n)$ is alternately smaller and larger than 1; since this fraction converges to $(\bar{x}-x)/(\bar{x}-x) = 1$, there exists n such that

$$0 < \frac{\bar{x} - c_{n-1}}{\bar{x} - c_n} < 1 . \qquad (17)$$

$q_{n-1} \leq q_n$, hence (16) and (17) give $-1 < \bar{m}_{n+1} < 0$.

Also $m_{n+1} > a_{n+1} \geq 1$, therefore m_{n+1} is reduced. //

Theorem 9.5. Let $x = [a_1, a_2, \ldots]$ be a quadratic irrational, and let the integers a, b, c, k, D, P_n, Q_n be as in Theorem 9.4.

Then :

$$P_{n+1} = a_n Q_n - P_n \qquad (n \geq 1) \qquad (18)$$

$$Q_n + 1 = \frac{D - P_{n+1}^2}{Q_n} \qquad (n \geq 1) \qquad (19)$$

$$a_{n+1} = \left[\frac{P_{n+1} + \sqrt{D}}{Q_{n+1}} \right] \qquad (n \geq 0) \qquad (20)$$

$$(-1)^n Q_1 P_{n+1} = D q_n q_{n-1} - (Q_1 p_n - P_1 q_n)(Q_1 p_{n-1} - P_1 q_{n-1}) \quad (n \geq 0) \qquad (21)$$

$$(-1)^n Q_1 Q_{n+1} = (Q_1 P_n - P_1 q_n)^2 - D q_n^2 \qquad (n \geq 0) \qquad (22)$$

$$Q_1 p_{n-1} = P_1 q_{n-1} + (q_{n-1} P_n + q_{n-2} Q_n) \qquad (n \geq 1) \qquad (23)$$

$$P_1 (q_{n-1} P_n + q_{n-2} Q_n) + D q_{n-1} = Q_1 (p_{n-1} P_n + p_{n-2} Q_n) \qquad (n \geq 1). \qquad (24)$$

Proof: (20) follows from the fact that $a_{n+1} = [m_{n+1}]$.

To prove (18) and (19) we note that

$$\frac{P_{n+1} + \sqrt{D}}{Q_{n+1}} = m_{n+1} = \frac{1}{m_n - a_n}$$

$$= \frac{Q_n}{(P_n + \sqrt{D}) - a_n Q_n} = \frac{-Q_n (a_n Q_n - P_n + \sqrt{D})}{(a_n Q_n - P_n)^2 - D} .$$

(21) and (22) are easily obtained by solving (15) for m_{n+1}, replacing x by $(P_1 + \sqrt{D}) / Q$, simplifying, and using $p_n q_{n-1} - p_{n-1} q_n = (-1)^n$.

(23) and (24) may be found from (15) by replacing n by $n-1$, x by

$(P_1 + \sqrt{D}) / Q_1$ and M_n by $(P_n + \sqrt{D}) / Q_n$, and crossmultiplying. //

With regard to the value of D, it should be noted that if $x = (A + \sqrt{N}) / B$ (A, B integers) , it is not always possible to take $D = N$. For example let $x = (1 + \sqrt{2}) / 2$ (which, incidentally, is reduced) ; then $a_1 = 1$, and $m_2 = 1/(x-a_1) = 2 + 2\sqrt{2}$ which is not of the form $(P + \sqrt{2})/Q$. As indicated in the statement of the theorems, D should be chosen by examining the quadratic equation satisfied by x. It is always possible to take $k = 1$, but sometimes $k = 2$ is useful, notably for $x = \sqrt{N}$.

Formulas (18) to (20) above provide a very convenient algorithm for expanding a quadratic irrational into a simple continued fraction : find D, get P_1 and Q_1 from $x = (P_1 + \sqrt{D}) / Q_1$, obtain $a_1 = [x]$, and then use (18) - (20) to compute P_2, Q_2, a_2, P_3, Q_3, a_3,... until some pair P_n, Q_n repeats a previous pair. Formula (22) is the basis for the solution of the Pell equation, as will be seen in later chapters.

Chapter 10. THE EXPANSION OF $\sqrt{D}$.

In the case where x is the square root of an integer, the simple-continued-fraction expansion has a particularly interesting form, which we now examine.

Theorem 10.1 Let D be a positive nonsquare integer, and let $x = \sqrt{D} = [a_1, a_2, \ldots]$, $m_n = [a_n, a_{n+1}, \ldots]$, $h = [\sqrt{D}]$. The expansion is of the form

$$\sqrt{D} = [h, \overline{b_1, b_2, \ldots, b_r, 2h}] , \tag{1}$$

where $0 \leq r \leq h(h+1) - 1$, $1 \leq b_i \leq h$, and $(b_1, b_2, \ldots, b_r)$ is symmetric, i.e. $b_i = b_{r+1-i}$. Each m_n is of the form

$$m_n = \frac{P_n + \sqrt{D}}{Q_n} \qquad (n \geq 1) \tag{2}$$

where P_n, Q_n are integers satisfying, for $n \geq 2$, $1 \leq P_n \leq h$, $1 \leq Q_n \leq 2h$. After $Q_1 = 1$, Q_{r+2} is the first Q_i to equal 1. We have the following formulas :

$$P_{n+1} = a_n Q_n - P_n \qquad (n \geq 1) \tag{3}$$

$$Q_{n+1} = \frac{D - P_{n+1}^2}{Q_n} \qquad (n \geq 1) \tag{4}$$

$$A_{n+1} = \left[\frac{P_{n+1} + h}{Q_{n+1}} \right] \qquad (n \geq 1) \tag{5}$$

$$(-1)^n P_{n+1} = Dq_n q_{n-1} - p_n p_{n-1} \qquad (n \geq 0) \tag{6}$$

$$(-1)^n Q_{n+1} = p_n^2 - D q_n^2 \qquad (n \geq 0) \tag{7}$$

$$p_{n-1} = q_{n-1} P_n + q_{n-2} Q_n \qquad (n \geq 1) \tag{8}$$

$$Dq_{n-1} = p_{n-1} P_n + p_{n-2} Q_n \qquad (n \geq 1) \tag{9}$$

Proof: x satisfies $x^2 - D = 0$, therefore taking $k = 2$ in Theorem 9.4, we have (2). $P_1 = 0$ and $Q_1 = 1$, so that (3) - (9) follow from (18) - (24) of Theorem 9.5 (with regard to (5), it is easily seen that if v is any real number and m, n are integers with $m > 0$, then

$$\left[\frac{n+v}{m}\right] = \left[\frac{n+[v]}{m}\right] ;$$

hence the $\sqrt{D}$ of formula (20), Chapter 9, can be replaced by h) . By lemma 1 (Chaper 9), $h + \sqrt{D}$ is reduced ; also $[h + \sqrt{D}] = 2h$, therefore by Theorem 9.3, $h + \sqrt{D} = [\overline{2h, b_1,\ldots,b_r}]$ $(r \geq 0)$. This can be rewritten as $[2h, \overline{b_1,\ldots,b_r, 2h}]$, hence $\sqrt{D} = [h,\overline{b_1,\ldots,b_r,2h}]$, proving (1). Taking $r + 1$ to be the fundamental period, Theorem 9.4 states that $r \leq h(h+1) - 1$. To prove symmetry, we have $-h + \sqrt{D} = [\overline{0,b_1,\ldots,b_r, 2h}]$, or, taking reciprocals, $-1 / (h - \sqrt{D}) = [\overline{b_1,\ldots,b_r,2h}]$; but by Theorem 9.2, $-1/(h-\sqrt{D}) = [b_r,\ldots,b_1,2h]$; since expansions are unique, we conclude that $(b_1,\ldots,b_r) = (b_r,\ldots,b_1)$, i.e. $(b_1,\ldots,b_r)$ is symmetric. Since (by Theorem 9.2) $m_2, m_3,\ldots$ are reduced, lemma 1 of Chapter 9 gives (for $n \geq 2$)

$$P_n < \sqrt{D} \tag{10}$$

$$Q_n - P_n < \sqrt{D} \tag{11}$$

$$P_n + Q_n > \sqrt{D} \tag{12}$$

Therefore $P_n \leq h$; also $Q_n - P_n \leq h$, hence adding gives $Q_n \leq 2h$; subtracting (10) and (12) gives $Q_n > 0$; subtracting (11) and (12) gives $P_n > 0$; thus $1 \leq P_n \leq h$ and $1 \leq Q_n \leq 2h$. Suppose $n \geq 2$ and $Q_n = 1$; then (10) and (12) give $\sqrt{D} - 1 < P_n < \sqrt{D}$, therefore

$P_n = h$; thus $m_n = h + \sqrt{D}$; but $r + 1$ is the fundamental period, hence, considering the expansion $h + \sqrt{D} = [\overline{2h, b_1, \ldots, b_r}]$, for which $m_{r+2} = m'_{r+2} = m'_1 = h + \sqrt{D}$ (primes refer to $h + \sqrt{D}$) , n is at least $r + 2$. Finally, this shows that $Q_j \geq 2$ $(2 \leq j \leq r+1)$, therefore by (5) , $a_j \leq (P_j + h) / Q_j \leq (h+h) / 2 = h$, i.e. $b_i \leq h$ $(1 \leq i \leq r)$. //

Examples : 1. $\sqrt{5} = [\, 2, \overline{4}\,]$

2. $\sqrt{8} = [\, 2, \overline{1, 4}\,]$

3. $D = 13$, $h = 3$

n	1	2	3	4	5	6
P_n	0	3	1	2	1	3
Q_n	1	4	3	3	4	1
a_n	3	1	1	1	1	6

$\sqrt{13} = [\overline{3, 1, 1, 1, 1, 6}]$, $r = 4$

4. $D = 31$, $h = 5$

n	1	2	3	4	5	6	7	8	9
P_n	0	5	1	4	5	5	4	1	5
Q_n	1	6	5	3	2	3	5	6	1
a_n	5	1	1	3	5	3	1	1	10

$\sqrt{31} = [\overline{5, 1, 1, 3, 5, 3, 1, 1, 10}]$, $r = 7$

Theorem 10.2 Using the same notation as in Theorem 10.1, we have the following :

(i) $(P_2, P_3, \ldots, P_{r+2})$ and $(Q_1, Q_2, \ldots, Q_{r+2})$ are each symmetric.

(ii) If $Q_n = Q_{n-1}$ and $n \leq r + 2$, then $r = 2n - 4$.

(iii) If $P_n = P_{n-1}$ and $n \leq r + 2$, then $r = 2n - 5$.

(iv) If $Q_n = 2$ and $n < r + 2$, then $P_{n+1} = P_n$, hence $r = 2n - 3$.

(v) If $b_{i=h}$, then $i = (r+1) / 2$. (i.e. only a central term of the symmetric part can equal h.)

Proof: Let $y_1 \sim y_2$ mean $y_1 \overline{y_2} = -1$; one can check that

$$\frac{P + \sqrt{D}}{Q} \sim \frac{R + \sqrt{D}}{S} \Leftrightarrow P = R,\ QS = D - P^2 . \tag{13}$$

In particular, using (4) , we have that (for $i, j \geq 2$)

$$m_i \sim m_j \Leftrightarrow P_i = P_j ,\ Q_{i-1} = Q_j . \tag{14}$$

Now let $2 \leq i \leq r + 2$; using Theorem 9.2 and the symmetry of $(b_1,\ldots,b_r)$, $m_i = \overline{[b_{i-1},\ldots,b_r, 2_h, b_1,\ldots,b_{i-2}]} \sim$ $\overline{[b_{i-2},\ldots,b_i, 2h, b_r,\ldots,b_{i-1}]} = \overline{[b_{r+3-i},\ldots,b_r, 2h, b_1,\ldots b_{r+2-i}]}$ $= m_{r+4-i}$, therefore (14) clearly shows that $(P_2,\ldots,P_{r+2})$ and $(Q_1, Q_2,\ldots,Q_{r+2})$ are symmetric. Suppose $2 \leq n \leq r + 2$, and $Q_n = Q_{n-1}$; shows that $m_n \sim m_n$; but, as just shown $m_n \sim m_{r+4-n}$, therefore $m_n = m_{r+4-n}$; since $r + 1$ is the fundamental period, this shows that $n = r + 4 - n$, i.e. $r = 2n-4$. Similarly, if $2 \leq n \leq r+2$ and $P_n = P_{n-1}$, then $m_n \sim m_{n-1}$, hence $m_{n-1} = m_{r+4-n}$; now $n = 2$ is impossible (it would make m_1 equal to a later m_i) , therefore we conclude $n-1 = r+4-n$, i.e. $r = 2n-5$.

Assume $Q_n = 2$ and $n < r + 2$. Formula (3) gives

$$P_n + P_{n+1} = 2a_n . \tag{15}$$

The symmetry of $(P_2,\ldots,P_{r+2})$, $(Q_1,\ldots,Q_{r+2})$, and $(a2,\ldots,a_{r+1})$ gives $Pn_{+1} = P_{r+3-n}$, $Q_n = Q_{r+3-n}$, and $a_n = a_{r+3-n}$. Therefore :

$$\begin{aligned}(m_{r+3-n} - a_{r+3-n}) - (m_n - a_n) &= m_{r+3-n} - m_n \\ &= \frac{P_{n+1} + \sqrt{D}}{Q_n} - \frac{P_n + \sqrt{D}}{Q_n} \\ &= \frac{P_{n+1} - P_n}{Q_n} \ .\end{aligned}$$

But $0 < m_i - a_i < 1$, hence $|P_{n+1} - P_n| < Q_n = 2$.

(15) shows that $P_{n+1} - P_n$ is even, therefore $P_n = P_{n+1}$, and $r = 2n-3$ follows by (iii). Finally, Theorem 10.1 shows that $Q_n \neq 1$ $(2 \leq n \leq r+1)$, while $Q_n > 2$ implies (by (5)) $a_n \leq 2h/Q_n < h$; therefore if $b_i = h$, we must have $Q_{i+1} = 2$, hence $r = 2(i+1) - 3$, i.e. $i = (r+1) / 2$. //

Formulas (3) - (5) provide a practical method for expanding $\sqrt{D}$ as a simple continued fraction. Because P_n, Q_n, and a_n are bounded, and no irrational numbers appear in the formulas, this method is especially convenient for rapid automatic calculation. Furthermore, (i) - (iii) of Theorem 10.2 show that only about half of the period needs to be calculated.

When r is odd, (1) is said to have a <u>central term</u> (namely b_i, where $i = (r+1) /2)$. When r is even, (1) is said to have no central term.

Chapter 11. THE PELL EQUATION.

The Diophantine equation $x^2 - Dy^2 = N$, where D and N are given integers and x and y are unknowns, is known as Pell's equation (John Pell, 1611-1685), although Pell was not the first to consider it. It appears in the famous cattle problem of Archimedes (see [1] p.249), and the Hindus, as long ago as 800 A.D., apparently could solve various cases of the equation, but it remained for Lagrange (1736-1813) to give a complete and elegant analysis of it, about one hundred years after Fermat (1601-1665) proposed the problem to the English mathematicians of his day.

The Pell equation is important for several reasons. By means of various substitutions, the solution of the general quadratic Diophantine equation $ax^2 + bxy + cy^2 + dx + ey + f = 0$ can be made to depend upon the solution of Pell's equation. Knowledge of the structure of the set of units in the field extension of the rationals by $\sqrt{D}$ (D being a positive nonsquare integer ; $x = a + b\sqrt{D}$ is said to be a unit if x and $1/x$ satisfy a monic quadratic equation with integer coefficients) depends upon a thorough knowledge of Pell's equation for $N = \pm 1$ and ± 4. Other applications include the minimization of indefinite quadratic forms (see LeVeque [5] , Chapter 8).

We shall take D to be a positive nonsquare integer, and concentrate primarily on the case $N = \pm 1$. It will be seen that the simple-continued-fraction expansion of $\sqrt{D}$ conveniently furnishes all solutions, if any exist. (When D is negative or a square the solutions are finite in number and usually not difficult to find directly, especially when N

is small.) It should be noted that if $D = k^2E$, the solutions of $x^2 - Dy^2 = N$ follow immediately from the solutions of $x^2 - Ey^2 = N$; thus it is sufficient to consider only square-free D; however, it will be just as easy to treat the general case.

Theorem 11.1. Let D be a positive nonsquare integer, and let N satisfy $|N| < \sqrt{D}$. Then any positive (i.e. $x > 0$, $y > 0$) solution x, y of $x^2 - Dy^2 = N$ with $(x, y) = 1$, satisfies $x = p_n$, $y = q_n$ for some $n \geq 1$, where p_n, q_n refer to the simple-continued-fraction expansion of $\sqrt{D}$.

Proof: First assume $N \geq 0$. Dividing by y^2 and factoring,

$$\left(\frac{x}{y} - \sqrt{D}\right)\left(\frac{x}{y} + \sqrt{D}\right) = \frac{N}{y^2} .$$

Therefore $x/y \geq \sqrt{D}$, and we have

$$\left|\frac{x}{y} - \sqrt{D}\right| = \frac{N}{\left(\frac{x}{y} + \sqrt{D}\right) y^2} \leq \frac{N}{2\sqrt{D}\, y^2} < \frac{1}{2y^2} .$$

Therefore by Theorem 7.2, there exists $n \geq 1$ such that $x/y = p_n/q_n$, hence $x = p_n$, $y = q_n$. For the case $N < 0$, we use the following clever argument. $y^2 - Ex^2 = M$, where $E = 1/D$ and $M = -N/D$; now $M > 0$, and $-N = |N| < \sqrt{D}$ gives $M < \sqrt{E}$, therefore the same argument as above gives that y/x is a convergent of $1/\sqrt{D}$ (and not the first, which is 0); since $\sqrt{D} > 1$, Theorem 5.1 gives that x/y is a convergent of $\sqrt{D}$. //

Theorem 11.2. Let D be a positive nonsquare integer, assume the notation of Theorem 10.1 and consider the Pell equations:

$$x^2 - Dy^2 = 1 \qquad (1)$$

$$x^2 - Dy^2 = -1 \qquad (2)$$

(1) has infinitely many solutions; if r is odd (i.e. the expansion of $\sqrt{D}$ has a central term), all positive solutions of (1) are :

$$(p_n, q_n) \ , \ n = k\,(r+1) \ , \ k = 1,2,3,\ldots \ ;$$

if r is even (i.e. no central term), all positive solutions of (1) are:

$$(p_n, q_n) \ , \ n = k\,(r+1) \ , \ k = 2, 4, 6,\ldots$$

(2) is solvable if and only if r is even ; if r is even, all positive solutions of (2) are :

$$(p_n, q_n) \ , \ n = k\,(r+1) \ , \ k = 1,3,5,\ldots$$

Proof: $1 < \sqrt{D}$, and any solution of (1) or (2) is relatively prime, hence by Theorem 11.1, any positive solution of (1) or (2) is (p_n, q_n) for some $n \geq 1$. Formula (7) of Theorem 10.1 gives

$$(-1)^n Q_{n+1} = p_n^2 - D\,q_n^2 \quad (n \geq 1) \qquad (3)$$

Therefore (p_n, q_n) is a solution of (1) if and only if n is even and $Q_{n+1} = 1$. But by Theorem 10.1, $P_{n+1} = 1$ (where $n \geq 1$) is equivalent to $n = k\,(r+1)$, where $k \geq 1$. The assertions about (1) clearly follow. For (2), we first note that no solution can have $x = 0$ or $y = 0$. Now (3) shows that (p_n, q_n) is a solution of (2) if and only if n is odd and $Q_{n+1} = 1$, i.e. n odd and $n = k(r+1)$ $(k \geq 1)$. Therefore there are no solutions if r is odd, while for r even, the positive solutions are as stated in the theorem. //

The remainder of this chapter shows how all positive solutions of

(1) and (2) may be found once the least positive solution is known. First, let us clarify the notion of "least" positive solution by observing that if (x_1, y_1) and (x_2, y_2) are different positive solutions of $x^2 - Dy^2 = N$ (where $D > 0$), then either $x_1 < x_2$ and $y_1 < y_2$, or $x_2 < x_1$ and $y_2 < y_1$: if $x_1 = x_2$, then also $y_1 = y_2$, therefore assume $x_1 < x_2$; then $Dy_2^2 - Dy_1^2 = x_2^2 - x_1^2 > 0$, and hence $y_1 < y_2$. Secondly, given x_1 and y_1, an equation

$$x_n + y_n \sqrt{D} = (x_1 + y_1 \sqrt{D})^n \qquad (4)$$

uniquely determines the integers x_n and y_n, since $\sqrt{D}$ is irrational and we can equate terms after expanding the power; in fact, denoting $x_1 + y_1 \sqrt{D}$ by u and its conjugate $x_1 - y_1 \sqrt{D}$ by v, we have $x_n - y_n \sqrt{D} = \overline{u^n} = \overline{u}^n = v^n$, therefore $x_n = (u^n + v^n)/2$ and $y_n = (u^n - v^n) / (2\sqrt{D})$. For example, $x_2 = x_1^2 + Dy_1^2$ and $y_2 = 2x_1y_1$.

Lemma: Let D be a positive nonsquare integer, and for convenience denote $\sqrt{D}$ by α. Let $|e| = |f| = 1$, $x_1^2 - Dy_1^2 = e$, $s^2 - Dt^2 = f$ (x_1, y_1, s, t positive), $m \geq 1$, and suppose

$$(x_1 + y_1 \alpha)^m < s + t\alpha < (x_1 + y_1 \alpha)^{m+1} . \qquad (5)$$

Then there exists a positive solution (a,b) of $x^2 - Dy^2 = e^m f$, such that $a + b\alpha < x_1 + y_1 \alpha$.

Proof: We have $1/(x_1 + y_1 \alpha) = e\,(x_1 - y_1 \alpha)$, so that dividing (5) by $(x_1 + y_1 \alpha)^m$ gives

$$1 < e^m (s + t\alpha)(x_1 - y_1 \alpha)^m < x_1 + y_1 \alpha . \qquad (6)$$

Let the middle member of (6) be $a + b\alpha$. Then $a + b\alpha < x_1 + y_1\alpha$, and $a^2 - Db^2 = (a + b\alpha)(a - b\alpha) = e^{2m}(s^2 - Dt^2)(x_1^2 - Dy_1^2) = e^m f$. Now $1/(a + b\alpha) = |a - b\alpha|$, therefore (6) gives $|a - b\alpha| < |< a + b\alpha$. But $2a = (a + b\alpha) + (a - b\alpha)$ and $2b\alpha = (a + b\alpha) - (a - b\alpha)$, hence $2a > 0$, $2b\alpha > 0$, from which $a > 0$ and $b > 0$. //

Theorem 11.3. Let D be a positive nonsquare integer, $\alpha = \sqrt{D}$, and consider the Pell equations (1) and (2). If (x_1, y_1) is the least positive solution of (1), then all positive solutions are (x_n, y_n) determined by (4), where $n = 1, 2, 3, \ldots$ If (x_1, y_1) is a least positive solution of (2), then all positive solutions are (x_n, y_n) determined by (4), where $n = 1, 3, 5, \ldots$, and furthermore (x_2, y_2) is the least positive solution of (1).

Proof: If x_n, y_n are defined by (4), then $x_n^2 - Dy_n^2 = (x_n + y_n\alpha)(x_n - y_n\alpha) = (x_1 + y_1\alpha)^n (x_1 - y_1\alpha)^n = (x_1^2 - Dy_1^2)^n$. Therefore if (x_1, y_1) is a positive solution of (1), so are (x_n, y_n) for $n = 1, 2, 3, \ldots$ Also, if (x_1, y_1) is a positive solution of (2), then so are (x_n, y_n) for $n = 1, 3, 5, \ldots$, and furthermore (x_n, y_n) for $n = 2, 4, 6, \ldots$ are positive solutions of (1).

(i) Suppose (x_1, y_1) is the least positive solution of (1), and let (s, t) be any positive solution. Now $x_1 + y_1\alpha > 1$ and $s + t\alpha \geq x_1 + y_1\alpha$, therefore there exists $n \geq 1$ such that $(s, t) = (x_n, y_n)$, for otherwise (5) holds for some $m \geq 1$, and the lemma provides a positive solution (a, b) of (1) less than (x_1, y_1).

(ii) Suppose (x_1, y_1) is the least positive solution of (2), and that (s_1, t_1) is a positive solution of (1) less than (x_2, y_2). If $s_1 + t\alpha > x_1 + y_1\alpha$. define $(s,t) = (s_1,t_1)$; if not, we have $s_1 + t_1\alpha < x_1 + y_1\alpha$, and there is an r such that $(s_1 + t_1\alpha)^{r-1} < x_1 + y_1\alpha < (s_1 + t_1\alpha)^r$, in which case define $s + t\alpha = (s_1 + t_1\alpha)^r$. In either case, (s, t) is a positive solution of (1) satisfying (5) for $m = 1$, therefore the lemma gives a positive solution (a,b) of (2) less than (x_1, y_1). This contradiction proves that (x_2, y_2) is the least positive solution of (1).

(iii) Suppose (x_1, y_1) is the least positive solution of (2) , and (s, t) is a positive solution of (2) not among (x_1, y_1) , (x_3, y_3),... Then $s + t\alpha$ exceeds $x_1 + y_1\alpha$ and is not a power of $x_1 + y_1\alpha$, therefore (5) holds for some $m \geq 1$. The lemma gives a positive solution (a, b) of (2) less than (x_1, y_1) , or else of (1) less than (x_2, y_2) (since $x_1 + y_1\alpha < x_2 + y_2\alpha$), according as M is even or odd. The former is impossible by assumption, the latter by (ii). //

It may be noted that is (a, b) is a solution of $x^2 - Dy^2 = 1$ and (x_1, y_1) is a solution of $x^2 - Dy^2 = N$, then (x_n, y_n) given by

$$x_n + y_n \sqrt{D} = (x_1 + y_1 \sqrt{D}) (a + b \sqrt{D})^n$$

is also a solution of $x^2 - Dy^2 = N$. However, there is no assurance that all solutions can be obtained in this way from one known solution.

Chapter 12. THE SOLVABILITY OF $x^2 - Dy^2 = -1$.

Since the Pell equation $x^2 - Dy^2 = -1$ is solvable if and only if the simple-continued-fraction expansion of $\sqrt{D}$ has no central term (Theorem 11.2), there naturally arises the problem of characterizing those D for which $\sqrt{D}$ has no central term. As yet, there is no complete solution of this problem, but some important partial results are known, and this chapter is devoted to presenting these results. The proofs given here are based on the material in Perron [8], Chapter 3, Theorems 20-22.

Theorem 12.1. Let the expansion of $\sqrt{D}$ have no central term, and assume the notation of Theorem 10.1. Then

$$D = Q_{n+1}^2 + P_{n+1}^2 , \tag{1}$$

where $n = (r + 2)/2$, and $(P_{n+1}, P_{n+1}) = 1$. Therefore (see Chapter 5) D is a product (possibly zero factors, i.e. the product is 1) of primes of the form $4k+1$ or twice such a product.

Proof: Since r is even, symmetry of $(Q_1, Q_2, \ldots, Q_{r+2})$ (Theorem 10.2) gives $Q_n = Q_{n+1}$, from which (1) follows by formula (4) of Chapter 10. To show that P_{n+1} and Q_{n+1} are relatively prime, we note from (7) of Chapter 10 that

$$(-1)^n Q_{n+1} = p_n^2 - D q_n^2 \tag{2}$$

$$(-1)^{n-1} Q_n = p_{n-1}^2 - D q_{n-1}^2 . \tag{3}$$

Therefore if a prime p divides $Q_n = Q_{n+1}$ and P_{n+1}, then p divides D by (1), and p divides p_n and p_{n-1} by (2) and (3); but

$(p_n, p_{n-1}) = 1$, since $p_n q_{n-1} - p_{n-1} q_n = (-1)^n$ formula (5) of Chapter 2). //

The converse of the above theorem is false. For example $205 = 5.41$ and $34 = 2.17$ are sums of two relatively prime squares, but

$$\sqrt{205} = [\, 14, \overline{3, 6, 1, 4, 1, 6, 3, 28} \,]$$

and $$\sqrt{34} = [\, 5, \overline{1, 4, 1, 10} \,]$$

have central terms. However, we have the following results, leading up to the main theorem, Theorem 12.4.

Theorem 12.2. Let $D \geq 3$ be nonsquare. Of the three equations

$$x^2 - Dy^2 = -1 \qquad (4)$$

$$x^2 - Dy^2 = 2 \qquad (5)$$

$$x^2 - Dy^2 = -2 \qquad (6)$$

at most one has a solution.

Proof: Any square is convergent to 0 or 1 (mod 4) and to 0 or 1 (mod 3) ; hence for $D = 3$, (4) and (5) are not solvable (consider the terms modulo 4 and 3 respectively), while (6) is solvable : $1^2 - 3.1^2 = -2$. Therefore assume $D \geq 5$, hence $2 < \sqrt{D}$. Clearly no solution has $x = 0$ or $y = 0$, and also any solution has $(x,y) = 1$. Therefore, using the notation of Theorems 10.1 and 10.2, Theorem 11.1 gives that any solution x,y is of the form $|x| = p_n$, $|y| = q_n$ for some $n \geq 1$. Hence, if (5) or (6) is solvable, then formula (7) of Theorem 10.1 shows that $Q_{n+1} = 2$, and Theorem 10.2 (iv) shows that

$$n + 1 = \frac{r + 3}{2} + k\,(r+1) \qquad (7)$$

where $k \geq 0$. Therefore r is odd, and (4) is not solvable. Also, since $(-1)^n Q_{n+1} = 2$ for (5) and $(-1)^n Q_{n+1} = -2$ for (6), (7) shows that (5) is not solvable if $(r+3)/2$ is even, while (6) is not solvable if $(r+3)/2$ is odd. Therefore at most one of (4) - (6) is solvable. //

Theorem 12.3. Let nonsquare D be a power (first or higher) of an odd prime, or twice such a power. Then one and only one of (4) - (6) is solvable.

Proof: Let $D = p^{\alpha}$ or $2p^{\alpha}$, where p is an odd prime and $\alpha \geq 1$. Since the case $D = 3$ was treated in the proof of Theorem 12.2, it is sufficient to assume $D \geq 5$ (so that $2 < \sqrt{D}$) and show at least one of (4) - (6) is solvable. Suppose (4) is not solvable. Then r is odd, and by Theorem 10.2 and (3) of Theorem 10.1, $P_n = P_{n+1}$ and $2P_n = a_n Q_n$, where $n = (r+3)/2$. Thus (8) and (9) of Theorem 10.1 give :

$$2p_{n-1} = (\, q_{n-1}\, a_n + 2q_{n-2}\,)\, Q_n \qquad (8)$$

$$2Dq_{n-1} = (\, p_{n-1}\, a_n + 2p_{n-2}\,)\, Q_n \qquad (9)$$

Therefore $Q_n \mid 2p_{n-1},\ 2Dq_{n-1}$. Suppose $k \mid Q_n,\ q_{n-1}$; then $k \mid 2p_{n-1},\ q_{n-1}$; (i) if $k > 1$ is odd, then $k \mid p_{n-1},\ q_{n-1}$, which is false since $(p_{n-1},\ q_{n-1}) = 1$; (ii) if $k > 1$ is even, then $2 \mid Q_n,\ q_{n-1}$, hence by (8) $4 \mid 2p_{n-1}$, therefore $2 \mid p_{n-1},\ q_{n-1}$, which is false. Therefore $(Q_n,\ q_{n-1}) = 1$, hence $Q_n \mid 2p_{n-1},\ 2D$. Now (formula (7) of Theorem 10.1) $(-1)^{n-1} Q_n = p_{n-1}^2 - D\, q_{n-1}^2$, therefore

$$(-1)^{n-1}\, 4 = Q_n \left(\frac{2p_{n-1}}{Q_n} \right)^2 - \left(\frac{2D}{Q_n} \right) 2q_{n-1}^2 . \qquad (10)$$

It follows that $(Q_n , 2D/Q_n) = 1, 2,$ or 4 ; also $Q_n \neq 1$ (since $2 \leq n \leq r + 1$), and $2D = 2p^{\alpha}$ or $4p^{\alpha}$, therefore one finds that $Q_n = 2D, D, D/2, 4,$ or 2.

(A) Q_n cannot be $2D$ or D : if $Q_n = 2D$ or D, then $2P_n = a_n Q_n \geq Q_n \geq D$, hence $P_n \geq D/2$. But (Theorem 10.1) $P_n \leq h < \sqrt{D} < D/2$.

(B) Q_n cannot be $D/2$: if $Q_n = D/2$, then Q_n is odd, hence a_n even, and $P_n = (a_n/2)\, Q_n \geq Q_n = D/2$, leading to a contradiction as in (A).

(C) Q_n cannot be 4 : if $Q_n = 4$, then $2D/Q_n$ is odd, $4 = Q_n$ divides $2p_{n-1}$, hence $2/p_{n-1}$, so that q_{n-1} is odd, and each term of (10) except the last is divisible by 4, which is impossible.

The only remaining possibility is $Q_n = 2$. Then $(-1)^{n-1}\, 2 = p_{n-1}^2 - Dq_{n-1}^2$ (formula (7) of Theorem 10.1), so that (5) or (6) is solvable. //

Theorem 12.4. Let nonsquare D be a power (first or higher) of a prime of the form $4n+1$, or twice a power (first or higher) of a prime of the form $8n+5$. Then $x^2 - Dy^2 = -1$ is solvable.

Proof: In view of Theorem 12.3, it is sufficient to prove that (5) and (6) are not solvable.

(i) Let $D = p^{\alpha}$ ($\alpha \geq 1$, $p = 4n+1$) . Then $D \equiv 1 \pmod 4$. Therefore $x^2 - Dy^2 \equiv 0-0,\ 0-1,\ 1-0,$ or $1-1 \pmod 4$, i.e. $0, 1,$ or $3 \pmod 4$. But $\pm 2 \equiv 2 \pmod 4$, therefore (5) and (6) are not solvable.

(ii) Let $D = 2p^{\alpha}$ ($\alpha \geq 1$, $p = 8n+5$). Then $x^2 - Dy^2 = \pm 2$ implies $x^2 \equiv \pm 2 \pmod{p}$, therefore $\left(\frac{2}{p}\right) = 1$ or $\left(\frac{-2}{9}\right) = 1$. But, recalling the values of $\left(\frac{2}{p}\right)$ and $\left(\frac{-2}{p}\right)$ given in Chapter 5, this is impossibe for $p = 8n + 5$. //

Corollary: If p is a prime of the form $4n+1$, then $x^2 - py^2 = -1$ is solvable, therefore by Theorem 12.1 the representation of p as the sum of two squares can be found by expanding $\sqrt{p}$ as a simple continued fraction.

For example, let $p = 13$. From Example 3 after Theorem 10.1, we have $r = 4$, $(r + 2)/2 = 3$, therefore $p = Q_4^2 + P_4^2 = 3^2 + 2^2$.

This construction for expressing a prime $p = 4n + 1$ as the sum of two squares is attributed to Legendre (1808) (see [7], Appendix I).

The converse of Theorem 12.4 is false. For example, $x^2-5.17y^2 = -1$ and $x^2-2.41\, y^2 = -1$ are solvable, since

$$\sqrt{85} = [\, 9, \overline{4, 1, 1, 4, 18}\,]$$

and $$\sqrt{82} = [\, 9, \overline{18}\,]$$

have no central terms.

CONCLUSION

It is hoped that the preceding chapters have demonstrated the utility and elegance of the theory of continued fractions as a tool in approximation theory and the theory of numbers. For although new methods have recently been developed in the field of Diophantine approximations, continued fractions remain the basic stepping stone, while in elementary number theory they provide one of the very few direct methods.

There are, of course, many topics in this area which the present survey has not discussed. For example, continued fractions can be used to assist in factoring numbers (see [1] p.266), and no mention has been made of the beautiful subject of the geometry of numbers, which is closely related to continued fractions. There are many directions for further study. Hurwitz's Theorem (Chapter 7) is the first of a whole series of related theorems and problems. One could explore continued fractions themselves in greater detail by referring to such books as Perron [8]. Alternatively, there is the extension to analytic continued fractions.

Finally, there are two challenging problems introduced by the material presented in this survey, problems which may provide subjects for further research. We have shown that the length of period of the simple-continued-fraction expansion of a quadratic irrational does not exceed $h(h+1)$ (see Theorem 9.4). However, this bound appears to be quite crude. For example, it means that the period of $\sqrt{D}$ is less

than about D, while for $D \leq 1000$ the largest period is 60 (for $D = 919$ and $D = 991$), and most periods are much less than 60. Very little seems to be known about this topic. Secondly, as discussed in Chapter 12, characterization of those D for which $x^2 - Dy^2 = -1$ is solvable (i.e. those D for which the period of $\sqrt{D}$ is odd) is far from complete. Theorem 12.4 is a fairly deep result, but more inclusive results may be found.

BIBLIOGRAPHY

1. Beiler, A. H. : Recreations in the Theory of Numbers, 2nd ed. Dover, 1966, N.Y. Chapter 22.

2. Hardy, G.H. and Wright, E.M. : An Introduction to the Theory of Numbers, 4th ed. Oxford, 1960. Chapters 10, 11.

3. Hurwitz : "Ueber die angenäherte Darstellung der Irrationalzahlen durch rationale Brüche," Mathematische Annalen , 39(1891), 279-284.

4. Khintchine, A. Ya. : Continued Fractions , 3rd ed. Noordhoff, 1963, Groningen. Sections I, II.

5. LeVeque, W.J. : Topics in Number Theory, Vol. 1. Addison-Wesley, 1956, Reading, Mass. Chapter 8.

6. Niven, I. and Zuckerman, H.S. : An Introduction to the Theory of Numbers, 2nd ed. Wiley, 1966, N.Y. Chapter 7.

7. Olds, C. D. : Continued Fractions. Random House, 1963, N.Y.

8. Perron, O. : Die Lehre von den Kettenbrüchen, 2nd ed. Chelsea, 1929, N. Y. Chapters 1-3.

9. Vinogradov, I.M. : Elements of Number Theory. Dover, 1954, N.Y.

10. Vorob'ev, N. N. : Fibonacci Numbers. Blaisdell.