Little Mathematics Library

N. M. BESKIN

FASCINATING FRACTIONS

Mir Publishers · Moscow

FASCINATING FRACTIONS

ПОПУЛЯРНЫЕ ЛЕКЦИИ ПО МАТЕМАТИКЕ

Н. М. Бескин

ЗАМЕЧАТЕЛЬНЫЕ ДРОБИ

Издательство «Вышэйшая школа» Минск

LITTLE MATHEMATICS LIBRARY

N. M. Beskin

FASCINATING FRACTIONS

Translated from the Russian by
V. I. Kisin, Cand. Sc. (Phys. and Math.)

MIR PUBLISHERS
MOSCOW

First published 1986
Revised from the 1980 Russian edition

На английском языке

Contents

Preface

This booklet is intended for high-school students interested in mathematics. It is concerned with approximating real numbers by rational ones, which is one of the most captivating topics in arithmetic.

In the last decade, some young mathematicians, and not only young mathematicians, have shown a negligent attitude towards "classical" and "pure" mathematics in contrast with "modern" and "applied" mathematics. This stance is fully unjustified.

First, mathematics rests on a foundation of numerous classical theories, facts and findings which must be known to every mathematician. For instance, the theory of continued fractions, a part of classical pure mathematics, is widely used nowadays to calculate numerical values of functions by means of computers.

Second, while science develops, many of its theories become obsolete and "dry up", like some branches of a tree. Quite a few do, yes, but not all of them. There are theories which survived centuries (or even millenia) and still retained their significance.

Continued fractions represent one of the most perfect creations of 17-18th century mathematicians: Huygens, Euler, Lagrange, and Legendre. The properties of these fractions are really striking.

The following should be borne in mind when reading this booklet.

Topics easily understandable are presented in normal print, while those more difficult are given in small print. Proofs of some theorems given in small print may be omitted safely. These theorems will necessarily be taken for granted.

However, mathematics is not just reading for entertainment. A future mathematician as well as a physicist or an

engineer has to acquire skill in dealing with mathematical constructions and proofs. So take a pencil and a sheet of paper and study carefully the topics given in small print. You may succeed in simplifying some proofs or finding better ones.

The theory of continued fractions is vast. This booklet covers only its fundamentals, but it contains everything that may be useful for a layman interested in mathematics. Professional mathematicians need to know much more.

Nikolai Beskin

Chapter 1

Two Historical Puzzles

1.1. Archimedes' Puzzle

1.1.1. Archimedes' Number. Many people believe that only a distant journey, preferably to outer space or the ocean bottom, could enable them to meet anything extraordinary, for the everyday life is so familiar that can show up no unusual facets.

What a delusion it is! Our surroundings are full of puzzles which go unnoticed because they seem to be habitual.

This chapter tells us a story of two enigmatic, yet familiar, episodes from the history of mathematics.

High-school students the world over know from the course in geometry a symbol π which denotes the ratio of the circumference of a circle to its diameter.

The letter π is the first letter of the Greek word $\pi\varepsilon\rho\iota\varphi\varepsilon\rho\varepsilon\iota\alpha$ which means "circle". An English mathematician Jones was the first to introduce the symbol π in 1706. In 1736 Euler adopted this notation instead of the symbol p he previously used. Since then the symbol π has come into general use.

From the most ancient times mathematicians sought a value for the number π. Archimedes determined its approximate value as 22/7*. This fact is so well known that hardly anybody suspects that it conceals a mystery. Who ever asks

---

* Actually Archimedes gave a different formulation to this result in his book *On the Measurements of the Circle*. He determined for π its bounds: $3\frac{10}{71} < \pi < 3\frac{1}{7}$. To quote Archimedes, "The circumference of any circle equals three times the diameter plus an excess which is less than one seventh of the diameter but greater than $\frac{10}{71}$ of it."

Although the value of π is closest to $3\frac{10}{71}$ as compared to $3\frac{1}{7}$, the simpler value $3\frac{1}{7}$ is the one in general use.

why Archimedes chose a fraction with 7 for denominator? What would happen if π were approximated by a fraction with denominator 8?

This question proves to be of extreme interest.

1.1.2. Approximation. Mathematicians often encounter a problem of replacing an object (a number, a function, a figure, etc.) by some other object of the same nature, which is in some sense sufficiently near to, but simpler than, the original one. This replacing is called the *approximation*. In the general

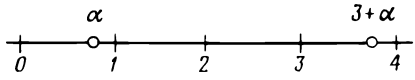

Fig. 1

case it requires that a set of objects be singled out and the sense of the phrase "sufficiently close to" be defined. We shall not discuss this general problem and restrict ourselves to the approximation of real numbers.

Let us consider the set of all real numbers. The conventional notation for this set is **R**. Real numbers may be of complicated nature, e.g. irrational numbers, or be cumbersome, e.g. fractions with large denominators.

It is worth explaining why cumbersomeness of a fraction is evaluated by its denominator. (We remind that a fraction is a number $\frac{p}{q}$ where p and q are *integers*, and $q \neq 0$; therefore, $\frac{\sqrt{3}}{3}$ and $\frac{\pi}{2}$ are not fractions.) If we are mainly interested not in the magnitude of a real number α but in its arithmetic nature, we need to know the position of α between two consecutive integers n and $n+1$. The addition of an integer to the number α will not change the arithmetic nature of α (this statement does not hold for that branch of arithmetic which deals with integers). Figure 1 shows two numbers α and $3+\alpha$ identically located within segments [0, 1] and [3, 4] (the term "segment" is defined on p. 41). For instance, the numbers $\frac{391}{4} = 97\frac{3}{4}$ and $\frac{3}{4}$ are identically located within the corresponding segments [97, 98] and [0, 1], and thus there are no reasons to regard the former as being more complicated than the latter. This implies that an analysis of the nature of the numbers within the segment [0, 1] would be quite sufficient since the same pattern is reproduced within each seg-

ment $[n, n+1]$. This is why we are concerned only with the denominator when evaluating the cumbersomeness of a fraction.

Let us single out a subset of fractions with a given denominator q from the set **R** of all real numbers. The distance between a number α and a fraction $\frac{p}{q}$ is $\left|\alpha - \frac{p}{q}\right|$. Now we can give an interpretation of the problem of the approximation of real numbers as follows: *to approximate a real number α by a fraction with q denominator which is the closest to α among all fractions with q denominator.*

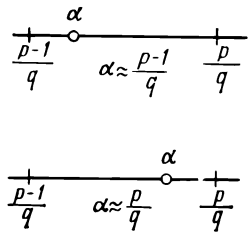

Fig. 2

If we mark all fractions with q denominator on the number line, the number α will fall within an interval between two fractions or coincide with one of them. The latter case is trivial, and we can write that

$$\frac{p-1}{q} < \alpha < \frac{p}{q}.$$

Of these two fractions the one nearest to α is chosen as its approximation (Fig. 2).

It could happen that α is the middle point of the segment $\left[\frac{p-1}{q}, \frac{p}{q}\right]$. This and only this case implies that there exist two solutions of the problem. *For the sake of definiteness*, we choose to adopt the left-end point of the segment as the approximation of α.

It is clear, therefore, that a fraction with any denominator can approximate the number α, that is, the choice of q denominator is a matter of preference.

Approximation is employed when it is necessary to use a rational number instead of an irrational number. It is also applicable to replacement of rational numbers by less cumbersome ones, i.e. by numbers with smaller denominators. For instance, the approximation of the number $\frac{2936}{7043}$ by the fraction with denominator 12 is

$$\frac{2936}{7043} \simeq \frac{5}{12},$$

since

$$\frac{5}{12} < \frac{2936}{7043} < \frac{6}{12},$$

where $\frac{2936}{7043}$ is nearer to $\frac{5}{12}$ than to $\frac{6}{12}$.

The approximation of real numbers by decimal fractions has long been in general use. However, decimals were yet unknown in Archimedes' time*, and he could choose fractions with arbitrary denominators to approximate the number π. Why did he prefer fractions with denominator 7? Could it be purely accidental?

1.1.3. Error of Approximation. A real number α is approximated by a fraction $\frac{p}{q}$ with an *error*

$$\Delta = \alpha - \frac{\hat{p}}{q},$$

where $\frac{\hat{p}}{q}$ stands for the end point of the segment $\left[\frac{p-1}{q}, \frac{p}{q}\right]$ which is the closest to α.

The error is thus the exact value of α minus its approximation.

Therefore, the error is positive if $\frac{\hat{p}}{q} = \frac{p}{q}$, and negative if $\frac{\hat{p}}{q} = \frac{p-1}{q}$.

The absolute value $|\Delta|$ of the error is called the *absolute error.*

It is clear that the absolute error does not exceed $\frac{1}{2q}$ (see Fig. 2):

$$|\Delta| \leqslant \frac{1}{2q}.$$

The number $\frac{1}{2q}$ is the *upper bound of absolute error.* The upper bound depends on the choice of approximation. For

---

* Decimal fractions became known in Europe at the end of the 16th century, although in the Orient they had been used since the end of the 15th century. They were invented by the Flemish scientist Simon Stevin. Here is what the English writer Jerome K. Jerome had to say on the matter: "From Gent we went to Bruges (where I had the satisfaction of throwing a stone at the statue of Simon Stevin, who added to the miseries of my school-days by inventing decimals), and from Bruges we came on here." (*Diary of a Pilgrimage*, the entry for Monday, June 9.)

instance, *if we agreed* to approximate the number α by the left-end point of the segment $\left[\frac{p-1}{q}, \frac{p}{q}\right]$, then the upper bound would be $\frac{1}{q}$.

1.1.4. Quality of Approximation. The absolute error approaches the upper bound if α is the middle point of the segment $\left[\frac{p-1}{q}, \frac{p}{q}\right]$. This is the most unfavourable case. If, however, α is very close to one of the end points, the actual absolute error may be considerably smaller than the upper bound.

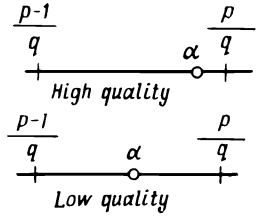

Fig. 3

This observation suggests that the evaluation of the *quality* of approximation is required. It is clear that *the approximation of a number α by a fraction with a small denominator is appropriate if the error is small*; or, to be more precise, if the absolute error is substantially less than the upper bound of the error (Fig. 3).

In order to evaluate the quality of approximation, we have to estimate the ratio of the actual absolute error to the upper bound on the absolute error

$$\frac{\text{absolute error}}{\text{upper bound on absolute error}} = \frac{|\alpha - p/q|}{1/2q} = 2\,|q\alpha - p|.$$

It is convenient to consider one half of this ratio denoted by h and called the *normalized error*,

$$h = |q\alpha - p|. \tag{1}$$

The normalized error h is thus one half of the ratio of the actual absolute error to the maximum possible error. It is obvious that

$$0 < h \leqslant \frac{1}{2}.$$

The quality of approximation is the higher, the less h is.

We call the quantity

$$\lambda = \frac{1}{2h} = \frac{1}{2\,|q\alpha - p|} \tag{2}$$

the *quality factor*. It has a simple and lucid meaning: *The quality factor of approximation is the factor by which the actual absolute error is less than the maximum possible error.* It is ob-

vious that

$$1 \leqslant \lambda < \infty,$$

and the greater λ, the better the approximation.

It would be wrong to expect fractions with greater denominators to be more useful. It could happen that the approximation of the number α by a fraction with the denominator 8 is less accurate than that by a fraction with denominator 7. Let us have a look at the number π, approximated by fractions with denominators from 1 through 10 (see Table 1). We omit the calculations, leaving them to the reader.

Table 1

q	Approximate value of π	Upper bound on absolute error	$\lvert\Delta\rvert$	h	λ
1	$\frac{3}{1}$	$\frac{1}{2}=0.5000$	0.1416	0.1416	3.5
2	$\frac{6}{2}$	$\frac{1}{4}=0.2500$	0.1416	0.2832	1.8
3	$\frac{9}{3}$	$\frac{1}{6}=0.1667$	0.1416	0.4248	1.2
4	$\frac{13}{4}$	$\frac{1}{8}=0.1250$	0.1084	0.4336	1.2
5	$\frac{16}{5}$	$\frac{1}{10}=0.1000$	0.0584	0.2920	1.7
6	$\frac{19}{6}$	$\frac{1}{12}=0.0833$	0.0251	0.1504	3.3
7	$\frac{22}{7}$	$\frac{1}{14}=0.0714$	0.0013	0.0089	56.5 (!)
8	$\frac{25}{8}$	$\frac{1}{16}=0.0625$	0.0166	0.1327	3.8
9	$\frac{28}{9}$	$\frac{1}{18}=0.0556$	0.0305	0.2743	1.8
10	$\frac{31}{10}$	$\frac{1}{20}=0.0500$	0.0416	0.4159	1.2

This table demonstrates that the approximation of π by fractions with denominator 7 is more accurate than that by the other fractions. The actual error is less than its upper bound by a factor of 56.5.

Figure 4 shows the location of π on the number line. Accidentally (but is it indeed accidental?) π happens to be quite

close to $3\frac{1}{7}$. If it were prescribed to approximate π with the absolute error less than or equal to 0.0013, how would we proceed? We would write down the condition

$$\frac{1}{2q} \leqslant 0.0013,$$

whence $q \geqslant 385$. Archimedes had achieved the same accuracy using a much smaller denominator. It is worth mentioning here that fractions with denominator 385 make it possible to approximate any real number with an error less than 0.0013, while fractions with denominator 7 are more preferable for approximating the π number.

Archimedes' choice could not therefore be accidental. But how did he make that choice?

Many centuries later (in 1585) a Dutch scientist from Metz, Adriaen Antoniszoon (also known as Adriaen Antonisz) found an approximate value $\frac{355}{113}$ for π.

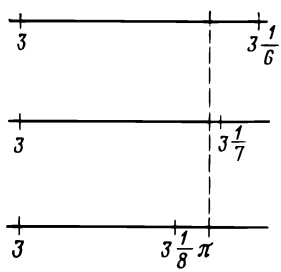

Fig. 4

This result has been published after Antoniszoon's death, by his son Adriaen Metius, so that the value 355/113 is traditionally called *Metius' number*. Metius' number has the same striking property as Archimedes' number: the actual error is less than it could be expected for the denominator 113. We invite the reader to examine Metius' number in the same way as Archimedes' number has been analysed.

There is no doubt that Metius' number was not an accidental discovery. In fact, it was known long before Adriaen Antoniszoon happened to find it (see, e.g. Struik's book in the Bibliography).

1.2. The Puzzle of Pope Gregory XIII

1.2.1. The Mathematical Problem of the Calendar. Pope Gregory XIII was not a mathematician but his name is associated with an important mathematical problem, that of the calendar.

Nature has supplied us with two obvious time units: the year and the day (solar day). We even read in one old text-

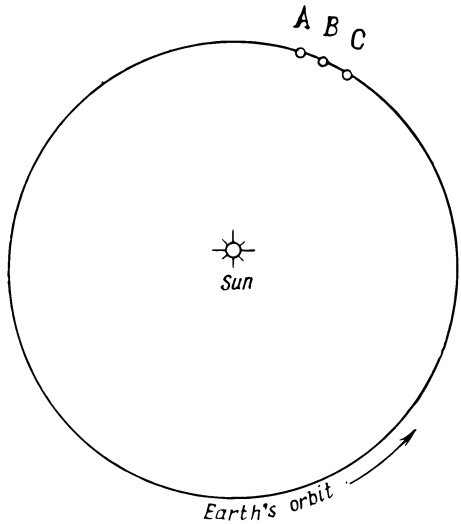

Fig. 5

book on cosmography: "*Unfotunately*, the year does not comprise an integral number of days." We could not but agree with this complaint because the fact does bring a lot of inconvenience. However, it also generates an interesting mathematical problem.

$$1 \text{ year} = 365 \text{ days } 5 \text{ hours } 48 \text{ minutes } 46 \text{ seconds}$$
$$= 365.242199 \text{ days}^*.$$

It would be impossible to enact and implement this duration of the year in civil life. But what if the civil year is declared exactly 365 days long? Figure 5 shows the orbit of the Earth. On January 1, 1985, at midnight, the Earth was at point A. On January 1, 1986, at midnight, it will be at point B, and next January 1 it will be at point C; and so forth. As a result, if we mark on the orbit the position of the Earth corresponding to a fixed date, this position will not be the same each year but will retard by nearly six hours.

---

* Neither the astronomical aspects of the calendar (such as variation in the length of the year) not its history are analysed here in detail; we concentrate only on one mathematical problem connected with the calendar. We recommend that the reader interested in these details look them up elsewhere.

The lag will build up to almost one day in four years, so that the fixed date will gradually slip among the seasons, that is, January 1 will move from winter first to autumn, then to summer. This would be inconvenient because periodic events such as crop sowing or the beginning of school year could not be tied to fixed calendar dates.

We know how to remedy the situation. Some years must be decreed to have 365 days, and some 366, in order to have the *average* duration of the year as close to the true duration as possible. This approach can approximate the true duration with any prescribed precision, but the required rule of alternation of shorter (ordinary) and longer (leap) years may be undesirably complicated. We need a compromise: a relatively simple pattern of alternation of the ordinary and leap years which brings the average length of the year sufficiently close to the true value.

1.2.2. Julian and Gregorian Calendars. This problem was first solved by Julius Caesar. Or rather, by the Alexandrian astronomer Sosigenes who was called for this purpose to Rome and given the job. The system introduced by Julius Caesar was as follows: three successive shorter (ordinary) years followed by one longer (leap) year. Much later, when the Christian chronology has been introduced, it was decided to have leap years when the number of the year was an integral multiple of 4.

This calendar is called the Julian calendar. The average length of the year in the Julian calendar is $365\frac{1}{4}$ days = 365 days 6 hours or 11 minutes 14 seconds longer than the true length.

The Julian calendar was improved by Pope Gregory XIII. In fact, calendar reforms had been proposed and elaborated long before but were never implemented. In 1582 he enacted the calendar reform. The alternation of ordinary and leap years was retained, with an additional rule: *If the number of the year ends with two zeros but the number of hundreds is not an integer multiple of 4, the year is treated as ordinary.* For example, this rule classifies the year 1700 as an ordinary and the year 1600 as a leap year. Furthermore, assuming that the error accumulated "since year 1 A.D." was 10 days, Pope Gregory XIII ordered to add 10 days to the current date, namely, to consider the day following the Thursday, October 4 of 1582, as Friday, October 15. There more days have been accumulated since then (in 1700, 1800, and 1900). Conse-

quently, at this moment the discrepancy between the Julian and Gregorian calendars is 13 days).

What is the average length of the Gregorian year? Of 400 years of the Julian calendar, 100 are leap years, while four Gregorian centuries contain only 97. Hence, the average Gregorian year has $365\frac{97}{400}$ days = 365.242500 days = 365 days 5 hours 49 minutes 12 seconds, or 26 seconds longer than the true length of the year.

We see that very high precision has been achieved by quite simple means. How could this result be achieved?

This question will be answered in Chapter VI.

Chapter 2

Formation of Continued Fractions

2.1. Expansion of a Real Number into a Continued Fraction

2.1.1. Algorithm of Expansion into a Continued Fraction. Let us forget for a time the decimal number system. The brilliant Soviet mathematician Nikolai Luzin (1883-1950) used to say in his lectures that "the advantages of the decimal system are zoological, not mathematical. If we had eight fingers on our two hands instead of ten, mankind would operate in the octal system." Decimal system is indeed very convenient in practice but it is inappropriate when theoretical aspects of arithmetic are discussed.

We thus forego the decimal and any positional number system, that is, we take Archimedes' place and ask ourselves: What would be the most natural approach to estimating a real number?

This question is answered without hesitation: the first step is to indicate the integers between which our number lies. For example,

$$\frac{61}{27} \text{ lies between 2 and 3,}$$

$$\sqrt{2} \text{ lies between 1 and 2,}$$

$$\pi \text{ lies between 3 and 4.}$$

Of course, it is sufficient to indicate only the lower of these bounds:

$$\frac{61}{27} = 2 + x \quad (0 < x < 1),$$

$$\sqrt{2} = 1 + y \quad (0 < y < 1),$$

$$\pi = 3 + z \quad (0 < z < 1).$$

Note that this estimation is not bound to any specific notation of integers, that is, to a specific number system.

Let us continue with the number $\frac{61}{27}$. Our estimate "two plus something" is too rough and constitutes only a *first approxi-*

mation. If we want to make the second step, we have to estimate the "makeweight" x. Since x is less than 1, it is logical to represent it by a fraction with numerator 1 (we again appeal to "acting natural", but we do it for the last time):

$$\frac{61}{27} = 2 + \frac{1}{x_1}.$$

Now x_1 is greater than unity, and we repeat the familiar steps: we single out the integral part of the number, and so forth. The reader is invited to follow attentively this alternation of steps:

$$\frac{61}{27} = 2 + \frac{7}{27} = 2 + \frac{1}{\frac{27}{7}} = 2 + \frac{1}{3 + \frac{6}{7}}$$

$$= 2 + \frac{1}{3 + \frac{1}{\frac{7}{6}}} = 2 + \frac{1}{3 + \frac{1}{1 + \frac{1}{6}}}.$$

The expression

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_{s-1} + \cfrac{1}{a_s}}}},$$

where $a_1, a_2, \ldots, a_s$ are *natural numbers** and a_0 is a natural number or nought, is called the *continued fraction.*

The numbers $a_0, a_1, a_2, \ldots, a_s$ are called the *terms* of the continued fraction. We can say that we have expanded the fraction $\frac{61}{27}$ into a continued fraction.

In what follows we shall often use this algorithm. It consists of two alternating steps.

Step 1. Single out the integral component of the number, that is, write it as the sum of an integer and a remainder less than unity.

Step 2. Represent the remainder as unity divided by a number greater than unity. Apply Step 1 to this denominator, and so on.

But before we go deeper into the theory of continued fractions, let us answer three questions.

* Recall that the natural numbers are 1, 2, 3, ... Nought is not included into the set of natural numbers.

2.1.2. Notation for Continued Fractions. *Question One.* Is not the notation for continued fractions too cumbersome? Our first example resulted in a three-storey fraction; for a twenty-storey fraction, the page space will not be enough.

This is why various notations have been devised for continued fractions. We shall use the following system

$$a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\ddots\cfrac{1}{a_s}}}=[a_0;\ a_1,\ a_2,\ \ldots,\ a_s].$$

The semicolon emphasizes the special role of the integral part a_0, distinct from that of the other terms. Its role is special but not more important; in this particular case it is rather less important.

2.1.3. Expansion of Negative Numbers into Continued Fractions. *Question Two.* How to expand a negative number into a continued fraction?

Two approaches can be used to expand a negative number into a continued fraction.

1. Place the minus sign in front of the fraction as a whole, for example,

$$-\frac{61}{27}=-\left(2+\cfrac{1}{3+\cfrac{1}{1+\cfrac{1}{6}}}\right)=-[2;\ 3,\ 1,\ 6].$$

2. Allow negative values of a_0, keeping $a_1, a_2, \ldots, a_s$ positive at all times. For example,

$$-\frac{61}{27}=-3+\frac{20}{27}=-3+\cfrac{1}{1+\cfrac{1}{2+\cfrac{1}{1+\cfrac{1}{6}}}}$$

$$=[-3;\ 1,\ 2,\ 1,\ 6].$$

In this book we shall use only the second approach. Hence, from now on a_0 *is an arbitrary integer while* $a_1, a_2, \ldots,$ *are natural numbers.*

Having made this remark, we shall not pay much attention to negative numbers when outlining the theory. A negative number can be obtained by adding a certain negative integer

to a positive number. In order to examine the arithmetical nature of the number $-\frac{61}{27}$, we can study the number $\frac{20}{27}$ and then add three unities to it.

2.1.4. Examples of Nonterminating Expansion. *Question Three.* Will the process of expanding a number α into a continued fraction inevitably terminate?

No, it may prove to be infinite. Let us have a look at some examples.

Example 1. Expand $\sqrt{2}$ into a continued fraction.

$$\sqrt{2}=1+\frac{1}{x_1};$$

$$x_1=\frac{1}{\sqrt{2}-1}=\sqrt{2}+1=2+\frac{1}{x_2};$$

$$x_2=\frac{1}{\sqrt{2}-1}.$$

We find that $x_2 = x_1$. Consequently, everything will repeat itself from this point on: $x_3 = x_2$, $x_4 = x_3$, ... We successively obtain

$$\sqrt{2}=1+\frac{1}{x_1}=1+\frac{1}{2+\frac{1}{x_2}}=1+\frac{1}{2+\frac{1}{2+\frac{1}{x_3}}}=\ldots$$

As long as we give a finite expression for $\sqrt{2}$ (involving an irrational x_n), we can use the equality sign. If this process is continued indefinitely, we obtain

$$\sqrt{2}\sim[1;\ 2,\ 2,\ 2,\ \ldots],$$

that is, the number $\sqrt{2}$ *corresponds* to a nonterminating continued fraction. We cannot put the equality sign between $\sqrt{2}$ and the nonterminating continued fraction [1; 2, 2, 2, . . .] because we are as yet unable to transform from one of these notations to another in *both* directions. So far the symbol of nonterminating continued fraction is devoid of meaning. We shall discuss and solve this problem in Chapter VI.

Example 2. In geometrical problems we can expand a geometrical quantity in a continued fraction without knowing the numerical value of the quantity. For example, let us find the ratio of the base to a leg of an isosceles triangle with the 108° vertex angle.

The angles of triangle ABC (Fig. 6) are 108°, 36°, 36°. We mark off $BB_1 = b$ (obviously, b can be marked off only once because $a < 2b$). We find

$$\frac{a}{b} = \frac{BC}{BB_1} = \frac{BB_1 + B_1C}{BB_1} = 1 + \frac{B_1C}{BB_1} = 1 + \frac{1}{x_1};$$

$$x_1 = \frac{BB_1}{B_1C} = \frac{AC}{B_1C}.$$

However, triangle B_1AC is similar to the initial triangle ABC. The first line above determined the ratio $\frac{a}{b}$ of the base to the leg. The second line represents the same problem be-

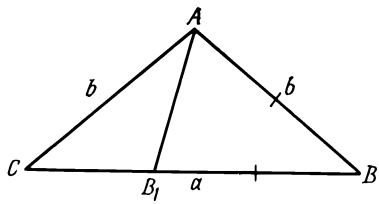

Fig. 6

cause x_1 is the ratio of the base to the leg in a triangle of the same shape. The process cannot terminate because the first step resulted in the reproduction of the initial situation. We can write

$$\frac{a}{b} \sim [1;\ 1,\ 1,\ 1,\ \ldots].$$

Likewise, it can be shown that

$$\frac{b}{a} \sim [0;\ 1,\ 1,\ 1,\ \ldots].$$

We shall return to this result at the end of Sec. 4.2.2.

Example 3. Expand the ratio of the diagonal of a square to its side into a continued fraction.

This example is more complicated than the second one. There we returned to the initial state after the first step, and here two steps are required.

If we assume that $\frac{d}{a} = \sqrt{2}$, this example coincides with Example 1. But the expansion of the ratio $\frac{d}{a}$ into a continued fraction can be obtained by geometrical means, without the numerical information.

Initial situation: we mark off the side along the diagonal, which can be done only once. We find (Fig. 7):

$$\frac{d}{a}=\frac{CA}{CB}=\frac{CB_1+B_1A}{CB}=1+\frac{1}{x_1};$$

$$x_1=\frac{CB}{B_1A}=\frac{AB}{AB_1}.$$

Now we erect $B_1B_2\perp AC$. Then $BB_2 = B_1B_2$ (please, prove it yourself). Now we supplement triangle AB_1B_2 to a square (merely for the sake of illustrative clarity; this is not necessary for the proof), and mark off AB_1 along BA. Having marked it off once, we obtain BB_2, the remainder being B_2A. Now we must mark off AB_1 along B_2A, but this is a repetition of the initial situation: the side of a square is marked off along the diagonal. Hence, the process is infinite, that is

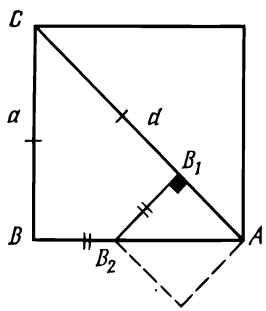

Fig. 7

$$x_1=2+\frac{1}{x_1};$$

$$\frac{d}{a}\sim[1;\ 2,\ 2,\ 2,\ \ldots].$$

It can be easily shown that

$$\frac{a}{d}\sim[0;\ 1,\ 2,\ 2,\ \ldots]$$

(see Subsec. 4.2.2).

2.2. Euclid's Algorithm

2.2.1. Euclid's Algorithm. The preceding Section dealt with the algorithm of the expansion of real numbers into continued fractions. This algorithm consisted of two alternating steps: (1) separation of the integral part of the number, and (2) presentation of the remainder (which is less than unity) as a reciprocal of a number greater than unity. This algorithm is a particular case of Euclid's algorithm which is widely used in mathematics.

Let us first illustrate how Euclid's algorithm operates in finding the greatest common divisor (abbreviated to GCD) of two natural numbers.

Let p and q be natural numbers. Euclid's algorithm consists of the following steps*:

Divide	Quotient	Remainder
p by q	a_0	r_0
q by r_0	a_1	r_1
r_0 by r_1	a_2	r_2
...	...	...

The following formulas describe the process:

$$\left.\begin{aligned}
p &= a_0 q + r_0 && (0 < r_0 < q);\\
q &= a_1 r_0 + r_1 && (0 < r_1 < r_0);\\
r_0 &= a_2 r_1 + r_2 && (0 < r_2 < r_1);\\
r_1 &= a_3 r_2 + r_3 && (0 < r_3 < r_2);\\
&\ldots\ldots\ldots && \ldots\ldots\ldots\\
r_{s-3} &= a_{s-1} r_{s-2} + r_{s-1} && (0 < r_{s-1} < r_{s-2});\\
r_{s-2} &= a_s r_{s-1}.
\end{aligned}\right\} \qquad (3)$$

Clarification. If the dividend and the divisor are integers, the remainder may be zero. Why then do we exclude the equality sign on the left, that is, why do we write, e.g. $0 < r_1 < r_0$ instead of $0 \leqslant r_1 < r_0$? Because if it so happens that $r_i = 0$, this equality terminates the sequence. The *algorithm cannot but stop* because the remainders $r_0, r_1, r_2, \ldots$ are non-negative integers, and each next r is less than the preceding one. Hence, the remainder will necessarily become zero at some step.

Equalities (3) can be rewritten as follows:

$$\begin{aligned}
\frac{p}{q} &= a_0 + \frac{r_0}{q};\\
\frac{q}{r_0} &= a_1 + \frac{r_1}{r_0};\\
\frac{r_0}{r_1} &= a_2 + \frac{r_2}{r_1};\\
&\ldots\ldots\ldots\\
\frac{r_{s-3}}{r_{s-2}} &= a_{s-1} + \frac{r_{s-1}}{r_{s-2}};\\
\frac{r_{s-2}}{r_{s-1}} &= a_s,
\end{aligned}$$

where r_{s-1} is what we seek: the GCD of p and q.

* If $p < q$, then $a_0 = 0$. This in no way disrupts the process.

Each of these equalities with the exception of the last one is an improper fraction presented as a sum of an integer and a proper fraction. Note that *the left-hand side of each equality (beginning with the second one) is the reciprocal of the proper fraction of the preceding equality.* We can, therefore, eliminate all r_i successively. By replacing the fraction $\frac{r_0}{q}$ in the first equality by its expression from the second equality, we find

$$\frac{p}{q}=a_0+\cfrac{1}{a_1+\cfrac{r_1}{r_0}}.$$

The fraction $\frac{r_1}{r_0}$ in the equality obtained above is now replaced with its expression from the third equality

$$\frac{p}{q}=a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\cfrac{r_2}{r_1}}}.$$

By continuing this process, we shall finally expand $\frac{p}{q}$ into a continued fraction. However, there is no need to repeat each time these substitutions. Indeed, $a_0, a_1, a_2, \ldots, a_s$ are the terms of the sought continued fraction. It only remains for us to remember the following rule:

In order to expand $\frac{p}{q}$ into a continued fraction, apply Euclid's algorithm to the numbers p and q. The quotients obtained in the successive divisions are the terms of the sought continued fraction.

Example. Expand the fraction $\frac{61}{27}$ into a continued fraction.

$$\frac{61}{7}\left|\frac{27}{2}\right. \quad \frac{27}{6}\left|\frac{7}{3}\right. \quad \frac{7}{1}\left|\frac{6}{1}\right. \quad \frac{6}{0}\left|\frac{1}{6}\right..$$

Hence, $\frac{61}{27}=[2;\ 3,\ 1,\ 6]$.

2.2.2. Examples of Application of Euclid's Algorithm. Euclid's algorithm can be used not only to find the GCD of two natural numbers. Let p and q be elements of an arbitrary set in which *division with remainder** has been defined. Euclid's algorithm can then be employed.

* This means that each ordered pair of elements p and q (p is the dividend and q the divisor) is put in correspondence with an ordered pair a and r (a is the quotient and r the remainder) which satisfies

For example, if p and q are segments of the number line, Euclid's algorithm can be applied to find their common measure. If p and q are commensurate, Euclid's algorithm terminates and the segment r_{s-1} [see formulas (3)] is their common measure. Indeed, it follows from the last equality in (3) that r_{s-1} is contained in r_{s-2} an integral number of times. By substituting r_{s-2} into the last-but-one equality, we obtain

$$r_{s-3} = a_{s-1}a_s r_{s-1} + r_{s-1} = (a_{s-1}a_s + 1)\, r_{s-1}.$$

Hence, r_{s-3} is also an integral multiple of r_{s-1}. If we thus climb the ladder of formulas (3) one step at a time, we reach the first two lines, that is, we prove that both p and q are integral multiples of r_{s-1}, and therefore r_{s-1} is the common measure of p and q.

In addition, Euclid's algorithm yields the terms of the continued fraction which corresponds to the ratio $\frac{p}{q}$. If the segments p and q are incommensurate, Euclid's algorithm does not stop, and the numbers a_0; a_1, a_2, . . . are the terms of the nonterminating continued fraction which represents the ratio $\frac{p}{q}$.

Euclid's algorithm is also applicable to polynomials of one variable x. The phrase "is less than" then means "is of lower power". By using the algorithm, we can find the GCD of two polynomials; however, this result has no bearing on our topic.

2.2.3. Summary. This chapter has outlined an algorithm (in its two versions) that permits to *expand* any real number α into a continued fraction, that is, to find a continued fraction *corresponding* to the number α.

If α is a rational number, it corresponds to a terminating continued fraction. In this case the calculations can be carried out in reverse order, that is, we can find the value of the continued fraction. For example,

$$2 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{6}}} = 2 + \cfrac{1}{3 + \cfrac{6}{7}} = 2 + \frac{7}{27} = \frac{61}{27}.$$

---

the conditions $p = aq + r$, $r < q$. Obviously, the operation of multiplication and the relation "is less than" must also be defined on this set.

Therefore, instead of the sentence "the continued fraction [2; 3, 1, 6] *corresponds* to the number $\frac{61}{27}$" we can say that "the number $\frac{61}{27}$ *equals* the continued fraction [2; 3, 1, 6]," or to be even more precise we can say that $\frac{61}{27}$ and [2; 3, 1, 6] are two different notations for the same number.

If, however, α is an irrational number, the situation is completely different. In this case the correspondence between α and the continued fraction is defined in one direction only: the number α *corresponds* to a nonterminating continued fraction, but not vice versa. We cannot *determine* a nonterminating continued fraction by the same procedure as used to calculate [2; 3, 1, 6]. So far we do not know the meaning of nonterminating continued fractions.

We are yet to solve this problem. It will be shown in Chapter 5 how to give the meaning to a nonterminating continued fraction. The reader will keep in mind, when reading Chapters 3 and 4, that so far we are unaware of that.

Chapter 3

Convergents

3.1. The Concept of Convergents

3.1.1. Preliminary Definition of Convergents. A continued fraction can be terminated by retaining the terms a_0; $a_1, a_2, \ldots, a_n$ and dropping all subsequent terms $a_{n+1}, a_{n+2}, \ldots$ The number obtained by this operation is called the *nth convergent* and denoted by $\frac{p_n}{q_n}$:

$$\frac{p_n}{q_n} = [a_0;\ a_1,\ a_2,\ \ldots,\ a_n] = a_0 + \cfrac{1}{a_1 + \ddots + \cfrac{1}{a_n}}.$$

Thus, for $n = 0$ we obtain the 0th convergent $\frac{p_0}{q_0} = [a_0] = \frac{a_0}{1}$.

Note 1. This is not the ultimate definition of a convergent. It will be defined in Subsec. 3.1.3.

Note 2. The concept of convergents is applicable both to terminating and nonterminating continued fractions. In the case of a terminating continued fraction we come to the last convergent coinciding with the continued fraction itself. For example, for the number $\frac{61}{27}$ we find

$$\frac{p_0}{q_0} = \frac{2}{1};$$

$$\frac{p_1}{q_1} = [2;\ 3] = \frac{7}{3};$$

$$\frac{p_2}{q_2} = [2;\ 3,\ 1] = \frac{9}{4};$$

$$\frac{p_3}{q_3} = [2;\ 3,\ 1,\ 6] = \frac{61}{27}.$$

If a continued fraction is nonterminating, the sequence of convergents is infinite. We do not know yet the meaning car-

ried by nonterminating continued fractions, but this fact does not present obstacles to understanding convergents. For example, for the fraction [1; 2, 2, 2, . . .] the sequence is

$$\frac{p_0}{q_0}=\frac{1}{1};$$

$$\frac{p_1}{q_1}=[1;\ 2]=\frac{3}{2};$$

$$\frac{p_2}{q_2}=[1;\ 2,\ 2]=\frac{7}{5};$$

. . . . . .

Hint. It is this fact (the possibility of forming the convergents) that enables us to breathe meaning into nonterminating continued fractions and assume that convergents are successive approximations which determine the value of the nonterminating continued fraction.

This hint is a seed from which the theory will grow. We develop this hint where we (in Chapter IV) prove that the convergents of terminating continued fractions are successive approximations, and for the time being we check it for the number $\frac{61}{27}$. In order to evaluate the approximations, note that $\frac{61}{27}\approx 2.259$.

Approximation		Error
number	value	
1	$\frac{2}{1}$	0.259
2	$\frac{7}{3}\approx 2.333$	−0.074
3	$\frac{9}{4}=2.250$	0.009

We note that the errors form a sequence with terms which have alternating signs and decreasing magnitude. Further on this pattern will be shown to constitute the general rule.

3.1.2. How to Generate Convergents. There is no need in writing out the whole multistorey continued fraction and carrying out the cumbersome process of successive evaluations

if we want to find the nth convergent. Quite simple recursion formulas* exist for evaluating p_n and q_n. Obviously,

$$\frac{p_0}{q_0} = \frac{a_0}{1};$$

$$\frac{p_1}{q_1} = a_0 + \frac{1}{a_1} = \frac{a_1 a_0 + 1}{a_1}.$$

In order to go from $\frac{p_1}{q_1}$ to $\frac{p_2}{q_2}$ we have to replace a_1 by $a_1 + \frac{1}{a_2}$. Elementary algebraic manipulations give

$$\frac{p_2}{q_2} = \frac{a_2(a_1 a_0 + 1) + a_0}{a_2 a_1 + 1}.$$

A careful look at this formula reveals the following structure:

$$\frac{p_2}{q_2} = \frac{p_1 a_2 + p_0}{q_1 a_2 + q_0}.$$

This formula reveals the general rule. Let us write it by giving separately the numerator and the denominator of the nth convergent:

$$\left.\begin{aligned} p_n &= p_{n-1} a_n + p_{n-2}, \\ q_n &= q_{n-1} a_n + q_{n-2}, \\ n &= 2, 3, \ldots, s. \end{aligned}\right\} \qquad (4)$$

Before proving formulas (4), let us first clarify their meaning. We shall not assign individual meanings to p_n and q_n (although this restriction will be dropped in the next Subsection). Formulas (4) are to be interpreted as follows: *we may take* p_n and q_n, as well as values proportional to them, as the numerator and denominator of the nth convergent. ▶ **Let us prove formulas (4) by mathematical induction. Assume that they hold for a certain fixed n that we denote by k

$$\left.\begin{aligned} p_k &= p_{k-1} a_k + p_{k-2}, \\ q_k &= q_{k-1} a_k + q_{k-2}. \end{aligned}\right\} \qquad (5)$$

and then prove that (4) hold for $n = k + 1$.

* A formula expressing an arbitrary element of sequence in terms of one or several preceding elements is said to be a recursion formula. Thus, the nth term of a geometric progression is given by a recursion formula $u_n = u_{n-1} q$ or by a non-recursion formula $u_n = u_1 q^{n-1}$. The recursion formula does not allow to evaluate u_n immediately; we have to find successively $u_1, u_2, \ldots, u_n$.

** The symbol ▶ marks the beginning of a proof.

An analysis of expressions

$$\frac{p_k}{q_k} = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_k}}},$$

$$\frac{p_{k+1}}{q_{k+1}} = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_k + \cfrac{1}{a_{k+1}}}}}$$

gives that *in order to go from* $\frac{p_k}{q_k}$ *to* $\frac{p_{k+1}}{q_{k+1}}$, *it is necessary to replace* a_k *by* $a_k + \frac{1}{a_{k+1}}$. Let us carry out this substitution in formulas (5). Note that p_{k-2}, q_{k-2}, p_{k-1}, and q_{k-1} remain unchanged because they do not contain a_k. We have

$$\begin{aligned} q_{k+1} &= p_{k-1}\left(a_k + \frac{1}{a_{k+1}}\right) + p_{k-2} \\ &= \frac{1}{a_{k+1}}\left[(p_{k-1}a_k + p_{k-2})\,a_{k+1} + p_{k-1}\right]; \\ p_{k+1} &= q_{k-1}\left(a_k + \frac{1}{a_{k+1}}\right) + q_{k-2} \\ &= \frac{1}{a_{k+1}}\left[(q_{k-1}a_k + q_{k-2})\,a_{k+1} + q_{k-1}\right]. \end{aligned}$$

Since p_{k+1} and q_{k+1} are defined to within a proportionality factor, we shall drop the factor $1/a_{k+1}$ and replace the expressions in parentheses via formulas (5):

$$\left.\begin{aligned} p_{k+1} &= p_k a_{k+1} + p_{k-1}; \\ q_{k+1} &= q_k a_{k+1} + q_{k-1}. \end{aligned}\right\}$$

We have therefore arrived at formulas (5) with $k + 1$ substituted for k.

Furthermore, we have already seen that formulas (4) hold for $n = 2$. We have thus proved that they also hold for $n = 2, 3, \ldots, s$. ■*

* The symbol ■ marks the end of a proof.

3.1.3. The Final Definition of Convergents. We are now ready to change the meaning of the term "convergent". We assume that the convergents of orders zero and one are the fractions $\frac{p_0}{q_0}$ and $\frac{p_1}{q_1}$, respectively, with $p_0 = a_0$, $q_0 = 1$, $p_1 = a_0a_1 + 1$, $q_1 = a_1$. Also, we assume that the convergents of orders 2, 3, . . ., s are the fractions whose numerators and denominators are given by formulas (4) for $n = 2, 3, \ldots, s$.

The reader may not have noticed the change. Did we really use a different interpretation of convergents?

The point is that the same number can be written in different ways. For example, notations 0.5, $\frac{1}{2}$, and $\frac{2}{4}$ stand for the same number. Until now the term "nth convergent" has been interpreted as a certain number regardless of notation.

Thus, different answers could be given in the Example of Subsection 2.1.2 to the question "What is the second convergent of $\frac{61}{27}$?": $2\frac{1}{4}$, 2.25, $\frac{9}{4}$, $\frac{18}{8}$, and so forth. All of them represent the same number in different notations. However, from now on *the term "convergent" will mean for us not only a definite number but also a prescribed notation of this number.* Thus, we decide that the convergent $\frac{p_2}{q_2}$ for $\frac{61}{27}$ is $\frac{9}{4}$ while $\frac{18}{8}$ would be an incorrect answer*. Now *both the numerator and denominator of each convergent are strictly defined, not to within a proportionality factor* (in the example above, $p_2 = 9$, $q_2 = 4$).

This convention is very important for the further elaboration of the theory of continued fractions.

If we note that all letters in formulas (4) stand for natural numbers, it will be easy to understand that *the denominators (as well as numerators) of successive convergents strictly increase*, that is, $q_n > q_{n-1}$, $p_n > p_{n-1}$ ($n = 2, 3, \ldots$). A comparison of p_0, q_0 with p_1, q_1 leads to

$$p_1 = p_0a_1 + 1, \quad q_0 = 1, \quad q_1 = a_1,$$

whence $p_1 > p_0$, while q_0 may prove to equal q_1. Finally,

$$\left.\begin{aligned} &q_0 \leqslant q_1 < q_2 < q_3 \ldots; \\ &p_0 < p_1 < p_2 < p_3 \ldots \end{aligned}\right\} \tag{6}$$

---

* Note, in passing, that $\frac{9}{4}$ and $\frac{18}{8}$ are indeed *different fractions*, even though they represent the *same rational number.*

Sequences (6) can be finite or infinite, depending on whether the continued fraction generating them is terminating or nonterminating.

3.1.4. Evaluation of Convergents. We give now a convenient way of arranging the results when evaluating convergents. The values of a_i will be written in the first, p_i in the second, and q_i in the third rows.

a_0	a_1	a_2	a_3			a_{s-1}	a_s
p_0	p_1	p_2	p_3			p_{s-1}	p_s
q_0	q_1	q_2	q_3			q_{s-1}	q_s

We begin with filling in the first row and the first two columns. The new entries are evaluated in the following order:

a_{n-2}	a_{n-1}	a_n
p_{n-2}	p_{n-1}	
q_{n-2}	q_{n-1}	

(1) the column $\begin{vmatrix} p_{n-1} \\ q_{n-1} \end{vmatrix}$ is multiplied by a_n; (2) the obtained column is added to the preceding one.

The same scheme is recommended if we want to calculate the value of a terminating continued fraction: the answer is given by the last column $\begin{vmatrix} p_s \\ q_s \end{vmatrix}$. This is much simpler than the step-by-step procedure.

The reader can practice filling in the table for a continued fraction [0; 3, 14, 1, 2, 5].

0	3	14	1	2	5
0	1	14	15	44	235
1	3	43	46	135	721

3.1.5. Complete Quotients. It is often necessary to terminate the process of expanding a number into a continued fraction before the end is reached. For example,

$$\frac{61}{27} = 2 + \frac{1}{\frac{27}{7}}$$

or

$$\frac{61}{27}=2+\cfrac{1}{3+\cfrac{1}{\cfrac{7}{6}}}.$$

The numbers $\frac{27}{7}$ and $\frac{7}{6}$ in the expressions above are called the *complete quotients* (the definition will be given below). The following notation is in use:

$$\frac{61}{27}=\left[2\,\middle|\,\frac{27}{7}\right]=\left[2;\ 3\,\middle|\,\frac{7}{6}\right]=[2;\ 3,\ 1,\ 6],$$

that is, the complete quotient is separated from the preceding terms by a vertical bar.

The complete quotient α_n can be defined in the following manner

$$\alpha=a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\ddots\cfrac{}{a_{n-1}+\cfrac{1}{\alpha_n}}}}, \tag{7}$$

where

$$\alpha_n=a_n+\cfrac{1}{a_{n+1}+\ddots}. \tag{8}$$

The complete quotient is thus a continued fraction which begins not with a_0 but with an arbitrary element a_n, that is, the continued fraction whose n terms, from a_0 to a_{n-1}, have been cut off. The symbolic notation of (7) is

$$\alpha=[a_0;\ a_1,\ a_2,\ \ldots,\ a_{n-1}\mid \alpha_n]. \tag{9}$$

Complete quotients possess the following property: *if some two complete quotients coincide, that is,* $\alpha_n=\alpha_{n+k}$ $(k>0)$, *then, first, this coincidence repeats itself after every* k *steps*:

$$\alpha_n=\alpha_{n+k}=\alpha_{n+2k}=\ldots=\alpha_{n+mk}=\ldots$$

and, second, the continued fraction itself is nonterminating and periodic (repeating).

The proof is obvious, and need not be given here. We shall only outline the first step. When we apply the algorithm of expanding a number α into a continued fraction and come to some complete quotient α_n, *our subsequent steps are independent of the preceding steps, that is, of the terms* $a_0,\ a_1,\ a_2,\ \ldots,\ a_{n-1}$. Therefore, the terms that follow a_{n+k-1} in the continued fraction are the same that follow a_{n-1}.

If α_n is a natural number, then $\alpha_n=a_n$, and the vertical bar in equality (9) can be replaced with a comma. It is logical to set $\alpha_0=\alpha$.

The continued fraction (8) can be a terminating or a nonterminating one. The meaning of nonterminating continued fractions will be

clarified in Chapter 4, so that for the time being we deal with (8) formally.

We are able now to derive a formula that relates complete quotients to convergents. A careful look at formula (7) reveals: if we drop $\frac{1}{\alpha_n}$ in the right-hand side, the remaining part of α is a convergent $\frac{p_{n-1}}{q_{n-1}}$ that can be rewritten, by using formulas (4), as follows:

$$\frac{p_{n-1}}{q_{n-1}} = \frac{p_{n-2}a_{n-1} + p_{n-3}}{q_{n-2}a_{n-1} + q_{n-3}}.$$

If we replace in this formula a_{n-1} by $a_{n-1} + \frac{1}{\alpha_n}$, we convert the left-hand side into α:

$$\alpha = \frac{p_{n-2}\left(a_{n-1} + \frac{1}{\alpha_n}\right) + p_{n-3}}{q_{n-2}\left(a_{n-1} + \frac{1}{\alpha_n}\right) + q_{n-3}} = \frac{(p_{n-2}a_{n-1} + p_{n-3})\,\alpha_n + p_{n-2}}{(q_{n-2}a_{n-1} + q_{n-3})\,\alpha_n + q_{n-2}}.$$

Finally,

$$\alpha = \frac{p_{n-1}\alpha_n + p_{n-2}}{q_{n-1}\alpha_n + q_{n-2}}.$$

3.2. The Properties of Convergents

3.2.1. The Difference Between Two Neighbouring Convergents. The step from the nth convergent to the $(n+1)$th is the *increment* of the nth convergent, denoted by Δ_n:

$$\Delta_n = \frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n} = \frac{p_{n+1}q_n - p_nq_{n+1}}{q_nq_{n+1}} = \frac{D_n}{q_nq_{n+1}}, \qquad (10)$$

where D_n denotes the numerator,

$$D_n = p_{n+1}q_n - p_nq_{n+1}. \qquad (11)$$

Let us reduce by 1 the subscripts of p_{n+1} and q_{n+1}, using formulas (4):

$$\begin{aligned} D_n &= (p_na_{n+1} + p_{n-1})\,q_n - p_n\,(q_na_{n+1} + q_{n-1}) \\ &= -(p_nq_{n-1} - p_{n-1}q_n). \end{aligned}$$

The expression in parentheses is of the same type as (11) but all subscripts are less by 1. Hence, it equals D_{n-1}:

$$D_n = -D_{n-1}.$$

This recursion relation enables us to reduce the subscript to zero:

$$D_n = -D_{n-1} = D_{n-2} = -D_{n-3} = \ldots = (-1)^nD_0.$$

The only step which separates us from total success is the direct evaluation of D_0:

$$D_0 = p_1q_0 - p_0q_1 = (a_1a_0 + 1)\cdot 1 - a_0a_1 = 1.$$

Hence,

$$D_n = p_{n+1}q_n - p_nq_{n+1} = (-1)^n \tag{12}$$

and by virtue of (10),

$$\Delta_n = \frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n} = \frac{(-1)^n}{q_nq_{n+1}}. \tag{13}$$

3.2.2. Comparison of Neighbouring Convergents. Let us mention some important properties of convergents.

Property 1. Each odd-numbered convergent is greater than the neighbouring convergents. Each even-numbered convergent is less than its neighbours.

When this regularity is tested, one must bear in mind that the zeroth convergent and the last one (provided it exists) have only one neighbour each.

Formula (13) immediately proves that Property 1 holds.

Property 1 indicates that successive convergents are alternately greater or less than their immediate predecessors.

Property 2. The difference between neighbouring convergents decreases in absolute value as the number of the convergent increases.

▶ Let us compare

$$|\Delta_n| = \frac{1}{q_nq_{n+1}};$$

$$|\Delta_{n+1}| = \frac{1}{q_{n+1}q_{n+2}}.$$

This gives $q_{n+2} > q_n$. Hence, the denominator of the second fraction is greater, and the fraction itself is smaller:

$$|\Delta_{n+1}| < |\Delta_n|. \blacksquare$$

Property 3. The exact value of a terminating continued fraction α is between any two neighbouring convergents. All even-numbered convergents lie to the left of α, that is, they give an approximation of α by defect. All odd-numbered convergents lie to the right of α, that is, they give an approximation of α by excess.

It is obvious that we have to exclude the last convergent which is exactly equal to α.

Instead of a formal proof, we shall only illustrate its main idea in Fig. 8.

Figure 8 shows how the convergents are arranged on the number line. The numeral marks not the magnitude but the number of each convergent. The leftmost point is the No. 0

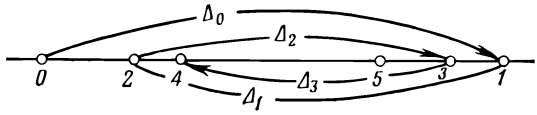

Fig. 8

convergent (i.e. the integral part of the continued fraction). A stop rightward is necessary to move from No. 0 to No. 1. This step (i.e. Δ_0) is shown by the upper arc. In order to move from convergent No. 1 to convergent No. 2 we need to step leftward, but this step (i.e. Δ_1) is shorter than Δ_0; then we make another step, and so on. We make steps alternately to the right and to the left, and each next step is shorter than the preceding one. Figure 8 convinces us that Property 3 holds.

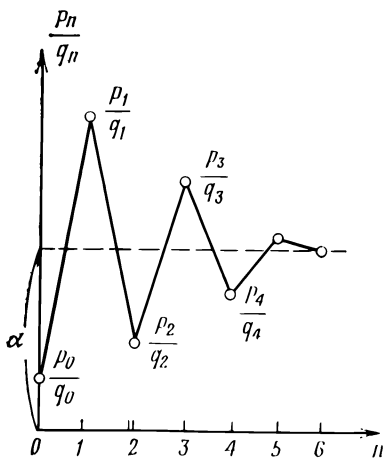

Fig. 9

Figure 9 is another illustration of the relative arrangement of the successive convergents. The abscissa axis gives the number of a convergent, and the ordinate axis presents its magnitude. The dashed line is drawn at the level of the value of α.

Property 4. The absolute error of approximating the number α *by a convergent* $\frac{p_n}{q_n}$ *is less than* $\frac{1}{q_n^2}$, *that is,*

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2}. \qquad (14)$$

▶ Indeed, by virtue of Property 3 and formula (13), we find

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}}.$$

This estimate is inconvenient because when approximating $\alpha \approx \frac{p_n}{q_n}$ we may not know the next convergent. For this reason we replace q_{n+1} in the last inequality by a smaller number q_n, thereby only strengthening the inequality. This proves inequality (14). ■

Property 4 shows that convergents are very good for approximating real numbers. If the fraction $\frac{p_n}{q_n}$ were not a convergent, the absolute error would be $\left|\alpha - \frac{p_n}{q_n}\right| \leqslant \frac{1}{2q_n}$.

3.2.3. Irreducibility of Convergents. Let us consider one more property of convergents.

Property 5. All convergents are irreducible to lower terms.

The numerators and denominators of convergents are given by formulas (4). Let us assume that the fraction $\frac{p_n}{q_n}$ can be reduced to lower terms, that is, its numerator and denominator have a common multiplier λ distinct from unity:

$$p_n = \lambda p'_n,$$
$$q_n = \lambda q'_n,$$

where p'_n and q'_n are natural numbers. Then formula (12) yields

$$\lambda (p_{n+1} q'_n - p'_n q_{n+1}) = (-1)^n.$$

This equality is absurd since the left-hand side is divisible by λ, while the right-hand side is not. Consequently, $\frac{p_n}{q_n}$ is not reducible to lower terms.

A set of mutually equal fractions contains only one irreducible fraction. The convergent may thus be defined as follows: *the convergent is the fraction, irreducible to lower terms, which gives the value of a truncated continued fraction.*

Chapter 4

Nonterminating Continued Fractions

4.1. Real Numbers

4.1.1. The Gulf Between the Finite and the Infinite. We can evaluate terminating continued fractions and hope that the reader does want to learn how to deal with nonterminating ones. It is precisely such desires that energize the scientific progress.

Any rational number can be converted to a terminating continued fraction. Conversely, any terminating continued fraction represents a rational number. Could it be that nonterminating continued fractions provide the means of representing irrational numbers?

Quite a few mathematical concepts that are familiar to us in finite form have captivating infinite analogues. Here are several examples.

The meaning carried by decimal fractions is quite clear. For example, 0.33 denotes $\frac{33}{100}$. And what is meant by 0.333...*?

The sum of a finite number of addends is also readily understandable. For example, $1+\frac{1}{2}+\frac{1}{4}=\frac{7}{4}$. But what about $1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots$?

There are finite polynomials, such as $1 + 2x + 3x^2$. But is it permissible to operate with "polynomials having an infinite number of terms", such as $1 + x + x^2 + \ldots + x^n + \ldots$?

In spite of the apparent similarity, the finite and infinite are separated by a deep and wide gulf. Mathematicians managed to overlook this gulf until the 19th century. Ignoring the danger, they treated infinite objects as they treated finite objects, and sometimes obtained absurd results. In the 19th

* The ellipsis is a mathematical symbol with two meanings. An ellipsis within a formula (e.g. $1 + x + x^2 + \ldots + x^n$) denotes a certain number of omitted terms; an ellipsis at the end (e.g. $1 + x + x^2 + \ldots$) stands for "and so on to infinity". Ellipses can also re place the whole rows.

century the way to deal with the infinite was gradually found, and reliable bridges were erected across the separating gulf. We shall walk across by one of these bridges.

Note that a terminating decimal fraction is in no way distinct from an ordinary fraction; the only distinction is the notation. The fraction 0.33 has the numerator 33 and the denominator 100. But what is the numerator of the nonterminating fraction 0.333...? We do not know the answer to this question, which is a clear indication that a nonterminating decimal fraction does not have the meaning carried by a terminating one. Academician N.N. Luzin used to say that drawing a symbol 0.333... does not impart a meaning to this symbol. It remains but a pattern. However, we can *give* this symbol a meaning.

The sum $1+\frac{1}{2}+\frac{1}{4}$ has a meaning because we can calculate it by successive additions: $1+\frac{1}{2}=\frac{3}{2}$, $\frac{3}{2}+\frac{1}{4}=\frac{7}{4}$. But we could not determine the infinite sum $1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots$ *by this method* because the process of consecutive additions will never terminate. And this is not a technicality but a principal obstacle. It would be frivolous to hide behind the argument that successive additions of terms in $1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots$ give us *approximate values* of the infinite sum. What does not exist, cannot be looked for. The meaning of infinite sums must *first* be defined, and only *after it* we can speak about the approximate values of the sum.

This is what we shall begin with now. Recall that we were going to cross the gulf between the finite and the infinite by one bridge (out of many). The name of this bridge is the *Principle of Nested Segments* or *Cantor's Continuity Axiom.**

4.1.2. Principle of Nested Segments. We often say that the number line is *continuous.* Mathematicians always have to seek logically impeccable statements that replace intuitive notions. The principle of nested segments is the axiom that expresses precisely the property of the number line that we call its *continuity.*

Recall that a *segment* is defined as a set of points of a number line, consisting of two distinct points a and b (called the

* Georg Cantor (1845-1918), the great German mathematician, created the set theory. The set theory became the foundation of the whole of mathematics.

ends of the segment) and all points between a and b. A segment is denoted by $[a, b]$. A set including all points between a and b but not the points a and b themselves is called the *interval* and is denoted by (a, b). The interval (a, b) contains two points less than the segment $[a, b]$, but in some cases this difference is extremely important. If one of the end points is added to the set of points between a and b, the result is a *half-closed interval.* The same letter can be used to denote a point of the number line and the number corresponding to it. Then

segment $[a, b] : a \leqslant x \leqslant b$;
interval $(a, b) : a < x < b$;
half-closed interval $[a, b) : a \leqslant x < b$;
half-closed interval $(a, b] : a < x \leqslant b$.

Let us consider on the number line a sequence of segments

$$[a_1, b_1], \quad [a_2, b_2], \ \ldots, \ [a_n, b_n], \ \ldots$$

having two properties: (1) each segment (beginning with the second one) is nested in the preceding segment, and (2) the length of the segments tends to zero (as $n \to \infty$).

The first property signifies that all points of the nth segment belong to the $(n - 1)$th segment (Fig. 10).

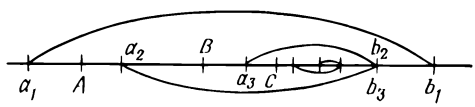

Fig. 10

The second property means that for an arbitrary fixed length ε there exists a number n such that the length of the segment $[a_n, b_n]$ is smaller than ε (and the segments with greater numbers are obviously even shorter).

In this case there exists a unique point that belongs to *all* segments.

Let us give a compact reformulation of the axiom.

Cantor's continuity axiom. *If an infinite sequence of segments is given on a straight line, such that (1) each next segment is nested within the preceding one, and (2) the length of the segments tends to zero, then there exists a unique point belonging to all these segments.*

Now we shall give a more detailed explanation of this axiom. Figure 10 shows the first several segments of our se-

quence. We define the nth step as the transition from the nth to the $(n+1)$th segment. Each step eliminates some of the points. For example, point A in Fig. 10 belongs to the first segment but not to the second one. Hence, this point will be eliminated at the first step of the process. Point B will survive the first step but gets eliminated at the second step. Point C survives the first two steps, but goes out on the third step, and so on. Each point of the segment $[a_1, b_1]$ has its own fate. Some fall inside the 1000th segment but stay outside of the 1001st. These points survive 1000 steps of the process but get eliminated at the 1001st step.

The principle of nested segments implies that there exists the point X which will *never* be eliminated, that is, it survives *each* step; in other words, it belongs to *any* segment, regardless of its number.

The axiom states that such point exists. As for the uniqueness of this point, it was introduced into the same formulation for the sake of convenience and can be readily proved. Indeed, let us assume that two such points exist, X and Y. We denote by d the distance between them. We have stipulated that the length of segments of the sequence tends to zero. Let us find a number n such that the length of segment $[a_n, b_n]$ is less than d,

$$|a_n, b_n| < d.$$

Then the segment $[a_n, b_n]$ cannot cover the segment $XY = d$, that is, points X and Y cannot belong to segment $[a_n, b_n]$ (and those following it). We have therefore proved that there cannot exist two points belonging to *all* segments.

Example 1. Let us consider the following segments on the number line

$$[0, 1], \left[\frac{1}{4}, \frac{3}{4}\right], \left[\frac{3}{8}, \frac{5}{8}\right], \left[\frac{7}{16}, \frac{9}{16}\right], \ldots,$$
$$\left[\frac{1}{2} - \frac{1}{2^n}, \frac{1}{2} + \frac{1}{2^n}\right], \ldots$$

Obviously, point $\frac{1}{2}$ and only it belongs to all these segments.

Example 2. Let a sequence of segments be given

$$[0, 1], \left[0, \frac{1}{2}\right], \left[0, \frac{1}{3}\right], \ldots, \left[0, \frac{1}{n}\right], \ldots.$$

Point 0 and only it belongs to all these segments.

In each of these examples we deal with a sequence of nested segments. We have easily singled out the unique point common to the segments. The principle of nested segments states that such a point *always* exists, no matter how the sequence was generated, provided that the two conditions are met.

Note. If we considered in Example 2 a sequence of *intervals*

$$(0,\ 1),\ \left(0,\ \frac{1}{2}\right),\ \left(0,\ \frac{1}{3}\right),\ \ldots,\ \left(0,\ \frac{1}{n}\right),\ \ldots$$

this sequence would not contain a point common to all intervals even though they are nested and their length tends to zero. Indeed, point 0 does not belong to any of them, while any other point of the interval (0, 1) will be left out at some step.

It is thus essential that Cantor's axiom be applied to *segments.* A similar statement would not hold for intervals.

The principle of nested segments expresses the continuity of a number line: segments converge to a point of the line, not to a "hole". Let us break the continuity of the line by piercing it at point $\frac{1}{2}$. To be precise, let us remove point $\frac{1}{2}$ from the number line. The remaining set of points, M, cannot be called a line. This is an assemblage of two so-called open rays, i.e. rays without vertices, $\left(-\infty, \frac{1}{2}\right)$ and $\left(\frac{1}{2},\ \infty\right)$. Let us follow Example 1 and consider a sequence of segments. Now these are not *segments of a line* because they lack one point, but segments on a set M. Each of them contains two ends and all the points of set M between them. Although these are nested segments and their length tends to zero, no point of set M can be found that belongs to all of them. The *principle of nested segments does not hold in M.*

4.1.3. The Set of Rational Numbers. Let us follow the process of gradual filling in of the number line with numbers. First we mark the integers. The set of all integers is traditionally denoted by **Z**. No subtle arguments are needed to show that the points of set **Z** do not fill up the number line completely. Each two integers are separated with a "solid" mass of points (an interval) that so far remain nameless.

Our next step is to mark rational numbers. It will be sufficient to mark all rational numbers within the interval between 0 and 1. All rational points on the number line will then be obtained by displacing the segment [0, 1] leftward and rightward an integral number of times.

We shall mark rational numbers on segment [0, 1] as follows:

Step 1. Mark off fractions with denominator 2. There is only one such a fraction: $\frac{1}{2}$.

Step 2. Mark off fractions with denominator 3, arranging them in the order of increasing numerators: $\frac{1}{3}$, $\frac{2}{3}$.

Step 3. Mark off fractions with denominator 4, arranging them in ascending order: $\frac{1}{4}$, $\left(\frac{2}{4}\right)$, $\frac{3}{4}$. The fraction $\frac{2}{4}$ is written in parentheses because this number has appeared earlier.

. . . . . .

Step $(n-1)$. Mark off fractions with denominator n, arranging them in ascending order: $\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n-1}{n}$. If fractions reducible to lower terms are encountered among those in this sequence, they can be crossed out.

. . . . . .

This process is infinite. Although we cannot complete it, we can be sure that it will mark *all* rational numbers in the interval between 0 and 1. Indeed, can there exist a fraction whose turn will never come? Let us select an arbitrary fraction from this interval, say, $\frac{37}{89}$. Obviously, at some step of marking fractions with denominators 2, 3, 4, . . . (namely, at the 88th step) we shall reach denominator 89. Then arranging the fractions in ascending order, $\frac{1}{89}, \frac{2}{89}, \frac{3}{89}$, and so forth, we inevitably reach $\frac{37}{89}$. Therefore, whatever fraction between 0 and 1 is selected, it will certainly be reached and marked off on the segment [0, 1]. Let us suppose that the process has been completed. This means that all rational points, i.e. those representing rational numbers, have been marked on the segment [0, 1]. By displacing these points by 1, 2, 3, . . . units leftward and rightward we shall have *all* rational numbers of the number line marked. In what follows we always denote the set of all rational numbers by **Q**.

4.1.4. The Existence of Nonrational Points on the Number Line. Is the number line completely filled up by the points of the set **Q**? No, it is not. Some points of the line do not belong to **Q**; they are not rational. However, this is not as obvious as for the set **Z**, and subtle arguments are needed to clarify this case. Pythagoras is reputed to have made the following

great discovery: there exists no number* whose square equals 2. An equivalent formulation is as follows: the diagonal of a square is incommensurate with its side. If all rational points are marked on the number line (Fig. 11), the arc of a circle whose radius is the diagonal of square OA passes freely through the numerical axis without intersecting the set **Q**.

Nevertheless, the set **Q** is *everywhere dense* on the number line. This means that *any* segment of the number line, however short, contains rational points. Consequently, even though

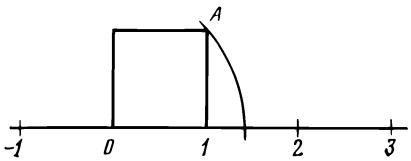

Fig. 11

rational points do not exhaust the number line, this line does not contain segments that would be absolutely free of rational points. This can be proved easily if the reader recalls how we went about marking the rational points on the segment [0, 1].

We shall consider sequences of nested segments

$$[a_1, b_1], \quad [a_2, b_2], \ \ldots, \ [a_n, b_n], \ \ldots$$

on the set **Q** (i.e. the ends of the segments are rational points). *The principle of nested segments does not hold on the set* **Q**. Even if the conditions of Cantor's continuity axiom are met, there may not exist a point in **Q** that belongs to all these segments. We shall immediately see that this fact can be used to invent numbers of a new type: irrational numbers.

4.1.5. Nonterminating Decimal Fractions. Let us attribute the following meaning to the symbol of a nonterminating decimal: a nonterminating decimal is a sequence of nested segments on the set **Q**. By terminating this fraction after each decimal place, we obtain the left ends of the segments. Adding unity to the last decimal place we find the right ends of the segments. For example, the fraction 0.313131... denotes the following sequence of nested segments on the set **Q**:

$$[0.3; 0.4], \ [0.31; 0.32], \ [0.313; 0.314], \ \ldots .$$

* No rational number, in fact. The qualification is not made because so far we do not know any other numbers.

Each step diminishes the length of these segments by a factor of 10, and hence, the length tends to zero, regardless of the fraction chosen.

Now we shall consider two examples that are superficially alike but in fact are profoundly different.

Example 1. A nonterminating periodical decimal 0.333... stands for the following sequence of nested segments:

$$[0.3;\ 0.4],\ [0.33;\ 0.34],\ [0.333;\ 0.334],\ \ldots$$

Is there a point belonging to all these segments? Here and below we mean a point in the set **Q**. No doubt, there is such a point on the number line.

This point is $x = \frac{1}{3}$.

The following inequalities hold:

$$\begin{aligned} 0.3 &< \tfrac{1}{3} < 0.4; \\ 0.33 &< \tfrac{1}{3} < 0.34; \\ 0.333 &< \tfrac{1}{3} < 0.334; \\ &\ldots\ldots\ldots\ldots \end{aligned} \tag{15}$$

For this reason the number $\frac{1}{3}$ is said to be the value of the nonterminating decimal 0.333... The formal definition of this concept will be given below, but first we shall discuss another example.

Example 2. Let us form two sequences: (a) the greatest decimal with 0, 1, 2, ... n, ... decimal places whose square is less than 2; and (b) the least decimal with 0, 1, 2, ..., n, ... decimal places whose square is greater than 2.

We successively find

$$\begin{aligned} 1^2 &< 2 \quad \text{but} \quad 2^2 > 2; \\ 1.4^2 &< 2 \quad \text{but} \quad 1.5^2 > 2; \\ 1.41^2 &< 2 \quad \text{but} \quad 1.42^2 > 2; \\ &\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

This process can be continued indefinitely. Is there a point in **Q** that belongs to all segments

$$[1;\ 2],\ [1.4;\ 1.5],\ [1.41;\ 1.42],\ \ldots\ ?$$

In other words, does there exist a rational number x that satisfies each of the following inequalities:

$$\begin{aligned} 1 &\leqslant x \leqslant 2, \\ 1.4 &\leqslant x \leqslant 1.5, \\ 1.41 &\leqslant x \leqslant 1.42, \\ &\ldots\ldots\ldots \end{aligned} \tag{16}$$

[Note that we use the sign $\leqslant$ (not $<$) because we are looking for a point belonging to the *segment*, that is, a point that may coincide with one of the ends. Accidentally, the number in Example 1 is always *within* a segment. If we began with a fraction 0.2000... corresponding to the number $\frac{1}{5}$, we would have to use the sign $\leqslant$.]

It is well known that there is no rational number that satisfied all the inequalities (16). This means that *any* rational number would violate inequalities (16), beginning with some line. It also means that the corresponding nonterminating decimal fraction 1.4142136... determined by the process described above is meaningless.

Now is the right moment to give it the meaning it lacked.

4.1.6. Irrational Numbers. The notation $\frac{1}{3}$ can be interpreted in two ways: (a) as a fraction $\frac{1}{3}$, that is, as a ratio of two natural numbers 1 and 3, or (b) as a nonterminating decimal 0.333..., that is, as the common point of nested segments

$$[0.3;\ 0.4],\ [0.33;\ 0.34],\ [0.333;\ 0.334],\ \ldots\ .$$

The first interpretation is inapplicable to the number x that we seek with inequalities (16). However, this number corresponds to a nonterminating decimal, that is, a system of nested segments

$$[1;\ 2],\ [1.4;\ 1.5],\ [1.41;\ 1.42],\ \ldots\ .$$

We can agree on a convention that this nonterminating decimal, or, what is the same, this system of nested segments, *defines* a number. This is a number of a new type: it cannot be presented as a ratio of natural numbers. Such numbers are called *irrational*.

Let us additionally clarify the idea of introducing irrational numbers.

The infinite sequence of nested segments (15) defines a number. It happens to be a rational number: $\frac{1}{3}$. We can operate with it ignoring the sequence (15).

The infinite sequence of nested segments (16) also defines a number but the type of this number is not familiar (we assume that only rational numbers were known) and the number appeared only as the sequence (15).

4.1.7. Real Numbers. The name for rational and irrational numbers is *real* numbers. In other words, the set of real numbers, **R**, is the union of the sets of rational and irrational numbers.

When the concept of number is extended and generalized, the old numbers should be treated as particular cases of a broader concept instead of being opposed to the new numbers. In other words, there must be a universal principle of formation and a universal notation for all real numbers.

The universal notation, also constituting the universal method of formation is that adopted for nonterminating decimals.

Some rational numbers can be presented as terminating decimals. However, we shall agree, in order to have a universal notation for *all* real numbers, to convert any decimal to a nonterminating decimal. This can be done in two ways, for example,

$$0.5 = 0.5000...,$$

$$0.5 = 0.4999....$$

In order to represent each real number by a nonterminating decimal in a unique manner, we agree on the following:

Convention. It is forbidden to use nonterminating decimals with the numeral 9 for the period.

Now the number 0.5 can be written as a nonterminating decimal in a unique way: 0.5000... .

By virtue of this convention, each real number is written as a nonterminating decimal fraction in a unique manner, that is, no two distinct nonterminating decimals can represent the same real number.

Let us emphasize that our definition actually identifies a real number as a nonterminating decimal fraction. Some real numbers can be written in other ways. For example, rational numbers are representable by common fractions. Roots from natural numbers are denoted by $\sqrt{2}$, $\sqrt{3}$, ..., $\sqrt{2}$, ... And finally, some numbers have been labelled by individual ("personal") symbols: π, e, and some others. However, non-

terminating decimals give a universal method of forming and presenting any real number.

This method of introduction of real numbers does not create the set **R** "from thin air". It assumes that a certain subset of **R**, namely, the set of all terminating decimals, already exists. The method allows us to *supplement* this set to **R** by using nested segments whose ends are given by terminating decimals. Real numbers can be defined differently, if we start not with the set of terminating decimals but with some other set which is *everywhere dense* on the number line.

The reader would be mistaken to think that we have already constructed the theory of real numbers. The definition of real numbers given above is only a first step. Many more steps would be needed to construct the theory, namely, ordering real numbers (i.e. finding a method of comparing the magnitudes), defining operations with real numbers (addition, multiplication, etc.), to name a few. However, we are not going to further elaborate this aspect. The purpose of this subsection is to clarify the principle of nested segments which we shall use to interpret nonterminating continued fractions.

4.1.8. Representing Real Numbers on the Number Line. Let a positive real number

$$x = \alpha_0, \ \alpha_1\alpha_2\alpha_3 \ \ldots \tag{17}$$

be given.

This is the decimal notation. Here α_0 is an arbitrary nonnegative integer, and the remaining α_i are numerals from 0 to 9. A terminating decimal fraction

$$\underline{x}_n = \alpha_0, \ \alpha_1\alpha_2 \ \ldots \ \alpha_n,$$

which is obtained if numerals beginning with α_{n+1} in (17) are dropped, is called the *approximate value of x with n decimal places by defect*. If we add one unity in the last decimal place,

$$\bar{x}_n = \alpha_0, \alpha_1\alpha_2 \ldots \alpha_n + 1 \cdot 10^{-n},$$

we obtain the *approximate value of x with n decimal places by excess*. If $\alpha_n = 9$ this unity can change the numerals preceding α_n. For example, for $x = \frac{892}{900} = 0.99111\ldots$, we have $\underline{x}_2 = 0.99$, $\bar{x}_2 = 1.00$.

Leaving aside the logical foundation, we can write the following apparent inequalities:

$$\underline{x}_n \leqslant x < \bar{x}_n.$$

Question to the reader. Why the sign in the left-hand inequality is $\leqslant$ while that on the right is $<$? Can they be reversed after a certain modification of the preceding definitions?

The following fact is of high importance: *if all real numbers are marked on the number line, the line is completely filled up.* We shall give now a better formulation of this statement and prove it.

Theorem 1. *Each real number corresponds to a unique point of the number line.*

▶ Let us take a positive real number $x = \alpha_0, \alpha_1\alpha_2\alpha_3 \ldots$. This number must belong to an infinite sequence of nested segments

$$[\underline{x}_0, \overline{x}_0], \quad [\underline{x}_1, \overline{x}_1], \quad [\underline{x}_2, \overline{x}_2], \quad \ldots \ .$$

The lengths of these segments form a geometric progression with common ratio $\frac{1}{10}$. By virtue of Cantor's continuity axiom, the number line contains a unique point belonging to all these segments. This is the point that corresponds to the number x. ■

Theorem 2. *Each point of the number line corresponds to a unique real number.*

▶ Let a point x be given on the number line (x lying somewhere on the right half-axis). If x is an integer, no more proof is needed. Otherwise, x lies between two neighbouring integers α_0 and $\alpha_0 + 1$. Let us begin the decimal notation of the number x with its integral part:

$$x = \alpha_0 \ldots$$

We divide the segment $[\alpha_0; \alpha_0 + 1]$ into ten equal parts. If x does not coincide with any one of the ten division points, it will be found between α_0, α_1 and $\alpha_0, \alpha_1 + 0.1$. We extend the decimal presentation of x:

$$x = \alpha_0, \alpha_1 \ldots$$

and divide the segment $[\alpha_0, \alpha_1; \alpha_0, \alpha_1 + 0.1]$ into ten equal parts.

If at some step of this process point x coincides with one of the division points, then

$$x = \alpha_0, \alpha_1\alpha_2 \ldots \alpha_n000 \ldots$$

If coincidence never occurs, then

$$x = \alpha_0, \alpha_1\alpha_2 \ldots \alpha_n \ldots$$

and x lies strictly inside all segments $[\underline{x}_n, \overline{x}_n]$ ($n = 0, 1, = 0, 1, 2, \ldots$). ■

Corollary. The set **R** *obeys the principle of nested segments.*

4.1.9. The Condition of Rationality of Nonterminating Decimals. We know from the course of high school mathematics that *each rational number can be expressed by a periodical decimal* (pure or mixed). For example,

$$\frac{1}{3} = 0.333 \ldots; \quad \frac{13}{90} = 0.1444 \ldots; \quad \frac{1}{5} = 0.2000 \ldots$$

Conversely, *each periodical decimal expresses a rational number.*

It then follows that *each irrational number is expressed by a nonperiodical nonterminating decimal.* For example, by using the algorithm of extracting square roots we can find any desired number of numerals in the decimal representing $\sqrt{2}$,

$$\sqrt{2} = 1.4142135 \ldots$$

We can always find one more numeral of the sequence. And though we do not know the formal law for the generation of this sequence of numerals (i.e. cannot find a function $\varphi(n)$ giving the nth decimal place), we are sure that this fraction is not periodical.

Conversely, *each nonperiodical decimal expresses an irrational.* For example, let us take a fraction

$$0.1010010001\ldots,$$

where the number of noughts between two consecutive unities each time increases by one. This fraction is nonperiodical and, therefore, it stands for an irrational number. In this example the formal law for the sequence of numerals is quite simple; if u_n is the nth decimal numeral, then

$$u_n = \begin{cases} 1 \text{ if } n \text{ is a number of the type } \frac{k(k+1)}{2}; \\ 0 \text{ otherwise} \end{cases}$$

4.2. Nonterminating Continued Fractions

4.2.1. Numerical Value of a Nonterminating Continued Fraction. By terminating the nonterminating continued fraction $[a_0; a_1, a_2, a_3, \ldots]$ after each successive term, we

generate the consecutive convergents:

$$\frac{p_0}{q_0}=[a_0],\quad \frac{p_1}{q_1}=[a_0;\ a_1],\quad \ldots,\quad \frac{p_n}{q_n}$$
$$=[a_0;\ a_1,\ a_2,\ \ldots,\ a_n],\ \ldots$$

And although a nonterminating continued fraction is only a symbol to which no numerical value is assigned, convergents are rational numbers. They define an infinite sequence of nested segments

$$\left[\frac{p_0}{q_0},\ \frac{p_1}{q_1}\right],\ \left[\frac{p_1}{q_1},\ \frac{p_2}{q_2}\right],\ \left[\frac{p_2}{q_2},\ \frac{p_3}{q_3}\right],\ \ldots,$$
$$\left[\frac{p_{n-1}}{q_{n-1}},\ \frac{p_n}{q_n}\right],\ \ldots \qquad (18)$$

We have already mentioned that the denominators of convergents strictly increase (see formulas (6)):

$$q_0 \leqslant q_1 < q_2 < q_3 < \ldots$$

All q_n being natural numbers, these inequalities mean that q_n grow indefinitely:

$$\lim_{n\to\infty} q_n = \infty .$$

But formula (13) then yields

$$\lim_{n\to\infty} \Delta n = \lim_{n\to\infty} \left(\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right) = 0.$$

The difference between neighbouring convergents tends to zero.

It is always assumed in such statements that $n \to \infty$.

Each segment (18) is nested within the preceding one (see Fig. 8). By virtue of Cantor's continuity axiom, there exists a unique point of the number line or, in other words, a unique real number, that belongs to all these segments. *It is this number that we define as the value of the nonterminating continued fraction.*

This definition implies the following corollaries:

1. *The value of a nonterminating continued fraction is between any two neighbouring convergents.*

2. *All even-numbered convergents are approximate values by defect, and all odd-numbered convergents are approximate values by excess, of the nonterminating continued fraction.*

Let us have a look at Fig. 9 once again. If the continued fraction is nonterminating, the broken line does not have the last segment whose end lies on the dashed horizontal line.

This broken line has an infinite number of vertices that alternately fall above and below the dashed line; α is the value of the nonterminating continued fraction.

3. *The sequence of even-numbered convergents*

$$\frac{p_0}{q_0}, \quad \frac{p_2}{q_2}, \quad \frac{p_4}{q_4}, \quad \ldots, \quad \frac{p_{2n}}{q_{2n}}, \quad \ldots$$

is monotone increasing and tending to α on the left. The sequence of odd-numbered convergents

$$\frac{p_1}{q_1}, \quad \frac{p_3}{q_3}, \quad \frac{p_5}{q_5}, \quad \ldots, \quad \frac{p_{2n+1}}{q_{2n+1}}, \quad \ldots$$

is monotone decreasing and tending to α on the right.

Let us look carefully at the sequence of segments (18). The ends of the segments are rational numbers. In our notation the first entry is alternately the left and the right end. This is obviously of no principal importance.

4.2.2. Representation of Irrationals by Nonterminating Continued Fractions. We have already learned the algorithm of expanding a real number into a continued fraction. Assume now that this algorithm is applied to an irrational α; what we obtain is a nonterminating continued fraction $[a_0; a_1, a_2, a_3, \ldots]$. Earlier this continued fraction was said to *correspond* to the number α:

$$\alpha \sim [a_0; a_1, a_2, a_3, \ldots].$$

Now we know how to determine the value of nonterminating continued fractions. A natural question then arises: is it α or another number that gives the numerical value of $[a_0; a_1, a_2, a_3, \ldots]$? In other words, is the correspondence between α and $[a_0; a_1, a_2, a_3, \ldots]$ symmetrical?

Yes, it is.

Indeed, the convergents obtained in the process of expanding α into a continued fraction are alternately greater and smaller than α. For example, consider the first two steps of the process:

$$\alpha = a_0 + \frac{1}{x_1},$$

whence

$$a_0 < \alpha.$$

Further,

$$x_1 = \frac{1}{\alpha - a_0} = a_1 + \frac{1}{x_2},$$

whence

$$\frac{1}{\alpha-a_0}>a_1,\quad \alpha-a_0<\frac{1}{a_1},\quad \alpha<a_0+\frac{1}{a_1}.$$

As a result,

$$a_0<\alpha<a_0+\frac{1}{a_1}$$

or, in a different form,

$$\frac{p_0}{q_0}<\alpha<\frac{p_1}{q_1}.$$

This chain of arguments can be prolonged further, that is, α is between any two neighbouring convergents.

If we wish to reverse the arguments and find the value of the obtained nonterminating continued fraction, we need to recall that by definition this value is the common point of all segments (18), that is, of all segments between neighbouring convergents.

However, there exists only one point that belongs to all segments (18). Therefore, the number α and the value of the continued fraction $[a_0; a_1, a_2, a_3, \ldots]$ coincide. We are thus entitled to replace the symbol $\approx$ with the equality sign $=$:

$$\alpha = [a_0; a_1, a_2, a_3, \ldots].$$

4.2.3. The Single-Valuedness of the Representation of a Real Number by a Continued Fraction. Do continued fractions offer a universal mean of representing real numbers? In other words, is it true that any real number* can be represented by a continued fraction, and in a unique manner?

The first part of the question has already been answered. Each real number can indeed be expanded into a continued fraction. A rational number expands into a terminating, and an irrational number to a nonterminating continued fraction. But the aspect of single-valuedness has not yet been analysed.

Let us think over the following example:

$$\frac{1}{6+\frac{1}{4}}=\frac{1}{6+\frac{1}{3+1}}=\frac{1}{6+\frac{1}{3+\frac{1}{1}}}$$

or, in contracted notation,

$$[0;\ 6,\ 4] = [0;\ 6,\ 3,\ 1].$$

* For the sake of simplicity, the arguments assume the numbers to be positive. It is clear, nevertheless, that the answer to the formulated question cannot change when considering negative numbers.

This transformation (the splitting-off of unity from the last term) can be effected in any fraction whose last term is distinct from unity. And if the last term is unity, it can be added to the last-but-one term (i.e. we can read the last example from right to left).

It can be readily proved that this is the *only* reason for the non-single-valuedness of the representation of a (positive) rational number by a continued fraction. We shall eliminate this cause by introducing the following convention: *The last term of a continued fraction must not be a unity.* From now it is *mandatory* for us to choose the first of two notations [0; 6, 4] and [0; 6, 3, 1] for the same number.

Now we are ready to prove that *two continued fractions* $[a_0; a_1, a_2, \ldots]$ *and* $[b_0; b_1, b_2, \ldots]$ (*terminating or nonterminating*) *are equal if and only if, first, they have identical numbers of terms and, second, their respective terms coincide, that is,* $a_0 = b_0$, $a_1 = b_1$, etc.

The condition "they have identical numbers of terms" must be understood as follows: either both fractions are terminating and have identical numbers of terms, or they are both nonterminating.

▶ Let us denote by α the value of two equal continued fractions (we do not know whether each of them is terminating or nonterminating):

$$\alpha = [a_0; a_1, a_2, \ldots] = [b_0; b_1, b_2, \ldots].$$

The term a_0 (as well as b_0) equals $E(\alpha)$* and thus splits off α in a single-valued manner. Consequently,

$$a_0 = b_0.$$

Subtract a_0 from α

$$\alpha - a_0 = [0; a_1, a_2, \ldots] = [0; b_1, b_2, \ldots]$$

and consider its reciprocal value

$$\frac{1}{\alpha - a_0} = [a_1; a_2, \ldots] = [b_1; b_2, \ldots].$$

---

* The definition of the function $E(\alpha)$ is: "The greatest integer not greater than α." For instance, $E\left(\frac{5}{2}\right) = 2$, $E(1) = 1$, $E\left(-\frac{5}{2}\right) = -3$. The symbol $E(\alpha)$ reads "the integral part of α"; the letter E comes from the French word "entier" (integral).

The term a_1 (and b_1) is $E\left(\frac{1}{\alpha - a_0}\right)$ and hence, is determined in a single-valued manner by the value of $\frac{1}{\alpha - a_0}$. Therefore,

$$a_1 = b_1,$$

and so forth. By repeating these arguments we shall prove that $a_2 = b_2$, $a_3 = b_3$, etc.

Can two equal continued fractions have unequal number of terms? Let us assume that the first continued fraction is a terminating one with s terms, while the second fraction is either a terminating one with t terms, $t > s$, or it is nonterminating. This means

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_s}}} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_s + \cfrac{1}{b_{s+1} + \ddots}}}}$$

or

$$a_s = a_s + \cfrac{1}{b_{s+1} + \ddots}$$

whence

$$\cfrac{1}{b_{s+1} + \ddots} = 0$$

which is impossible. Hence, $t = s$.

We conclude that each real number is expressed by a continued fraction, in a unique manner. ■

In this proof we have employed a rule that is often useful. When the expansion of α into a (terminating or nonterminating) continued fraction is known, the following should be done in order to find the expansion of $\frac{1}{\alpha}$:

(1) move the whole "comb" one step to the right, if $a_0 \neq 0$, and write in a nought instead of the integer, or

(2) if $a_0 = 0$, move the whole "comb" one step to the left.

Examples.

$$\alpha = [3;\ 1,\ 2,\ 5],\quad \frac{1}{\alpha} = [0;\ 3,\ 1,\ 2,\ 5];$$

$$\beta = [0;\ 2,\ 2,\ 2,\ \ldots],\quad \frac{1}{\beta} = [2;\ 2,\ 2,\ \ldots].$$

The proof follows from a careful examination of the following continued fractions:

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}} = [a_0;\ a_1,\ a_2,\ \ldots];$$

$$\frac{1}{\alpha} = \cfrac{1}{a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}}} = [0;\ a_0,\ a_1,\ a_2,\ \ldots].$$

4.3. The Nature of Numbers Given by Continued Fractions

4.3.1. Classification of Irrationals. We already know the following important fact: *each rational number is given by a terminating continued fraction, and irrational number by a nonterminating continued fraction.*

We shall not add anything here on rational numbers. But irrationals may be of very different types. Let us get acquainted with their classification.

The equation

$$a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n = 0, \qquad (19)$$

where $a_0 \neq 0$, is called the *algebraic equation of degree n.* We shall consider only the case when all coefficients in equation (19) are rational or even integral. These two cases are identical. If the coefficients are fractional, we can multiply both sides of the equation by the common denominator of these fractional coefficients, thus obtaining an equation with integral coefficients equivalent to the initial equation.

In our discussion of equation (19) we assume hereafter that its coefficients are integers (positive, negative, or zeros). An additional constraint is imposed on the coefficient of the leading term: $a_0 \neq 0$.

A real number is called an algebraic number of degree n if this number is a root of an algebraic equation of degree n with integral coefficients but is not a root of any other algebraic equation of lower degree with integral coefficients.

Example 1. Each rational number $\frac{p}{q}$ is an algebraic number of the first degree because it is the root of equation

$$qx - p = 0.$$

Example 2. The number $\sqrt{2}$ is an algebraic number of the second degree because it is a root of the equation

$$x^2 - 2 = 0.$$

We know that $\sqrt{2}$ cannot be a root of any equation of the first degree with integral coefficients because such an equation $(a_0x + a_1 = 0)$ has a rational number $x = -\frac{a_1}{a_0}$ for the root.

Algebraic numbers of the second degree are called *quadratic irrationalities.*

It was discovered that there exist nonalgebraic numbers. These numbers are called *transcendental numbers.*

Here is their definition: *A real number α is said to be transcedent if it is not a root of any algebraic equation with integral coefficients.*

To discover a transcendental number is not an easy task. If we want to prove that a number α is algebraic, it is sufficient to find an algebraic equation with integral coefficients for which α is a root. But if we cannot find this equation, we cannot conclude that α is transcendental, for we have to prove that no such equation exists. This problem was solved for the first time by the French mathematician Joseph Liouville in 1844. He proved the transcendency of some specific real numbers. In 1882 the German mathematician Ferdinand Lindemann proved that π is a transcendental number. At present very many examples of transcendental numbers are known. For example, decimal logarithms of all rational numbers, with the exception of numbers of the type 10^n, are transcendental.

The reader must be warned against a misunderstanding. The fact that examples of transcendental numbers are difficult to find does not mean at all that they are rare. Quite the opposite! Georg Cantor showed that *in a certain sense* (it would not be possible to explain what this means in this booklet) almost all real numbers are transcendental, that is, algebraic numbers are rare exceptions. However, the nature of algebraic numbers is simpler, and therefore we can give numerous

examples of them. As for transcendental numbers, it is always very hard to prove the transcendence in each specific case*.

4.3.2. Quadratic Irrationals. We know from the preceding subsection that a *quadratic irrational is an irrational number which is a root of a quadratic equation with integral coefficients.*

The word "irrational" replaces the phrase found in the preceding definition: "and it is not a root of any algebraic equation of lower degree with integral coefficients". In the case in question this means "is not the root of any equation of first degree with integral coefficients", that is, *is not a rational number.*

Let us consider a quadratic equation

$$a_0x^2 + a_1x + a_2 = 0,$$

where a_0, a_1, a_2 are integers and $a_0 \neq 0$. Its roots are given by the formula

$$x = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_0a_2}}{2a_0}.$$

These roots are quadratic irrationals if the following necessary and sufficient conditions are satisfied:

(1) the discriminant $D = a_1^2 - 4a_0a_2$ must be non-negative. If $D < 0$, the roots would not be real;

(2) the discriminant D must not be an exact square. If $D = N^2$, the roots would be rational.

These conditions enable us to give a different definition of the quadratic irrational: *a quadratic irrational is a number of a type* $p + q\sqrt{D}$ *where* p *and* q *are rational numbers and* D *is a natural number which is not an exact square.*

Before analysing some examples of quadratic irrationals, let us prove in advance four useful lemmas. But first we shall introduce some notations and define some terms that will help us to avoid repetitions.

Lower-case Roman letters p, q, . . . will always denote *rationals* (positive, negative, or zero). In particular cases they may happen to be integers.

Capital Roman letters D, M, N, . . . will denote *natural numbers not equal to exact squares*: 2, 3, 5, 6, 7, 8, 10, . . .

---

* Actually, a simple method is known for constructing continued fractions whose values are transcendental numbers. However, if one faces the problem of proving the transcendence of a number that has been defined by different means (π, log 2, sin 1, etc.), this always constitutes a very difficult problem.

Two square-root radicals* $\sqrt{M}$ and $\sqrt{N}$ are said to be *similar* if $\sqrt{N} = p\sqrt{M}$. Otherwise, that is, if the ratio $\frac{\sqrt{N}}{\sqrt{M}}$ is not rational, the radicals $\sqrt{M}$ and $\sqrt{N}$ are not similar. For example, $\sqrt{2}$ and $\sqrt{8}$ are similar but $\sqrt{2}$ and $\sqrt{10}$ are not.

If $\sqrt{M}$ and $\sqrt{N}$ are nonsimilar radicals, then $\sqrt{MN}$ is also a square-root radical (i.e. it is not an exact square) nonsimilar to either of the initial two. This is clear from the identities

$$\sqrt{MN} = N\frac{\sqrt{M}}{\sqrt{N}} \text{ (not a rational number)},$$

$$\frac{\sqrt{MN}}{\sqrt{M}} = \sqrt{N}, \quad \frac{\sqrt{MN}}{\sqrt{N}} = \sqrt{M}.$$

Lemma 1. *If $\sqrt{M}$ and $\sqrt{N}$ are nonsimilar radicals, then the equality*

$$k + l\sqrt{M} \quad m\sqrt{N} = 0 \tag{20}$$

holds only if $k = l = m = 0$.

In a more concise notation,

$$k + l\sqrt{M} + m\sqrt{N} = 0 \Leftrightarrow k = l = m = 0.$$

▶ Two cases must be considered in the proof: (1) $l \neq 0$ and $m \neq 0$ (for any k), and (2) one of the coefficients l, m is nonzero while the other vanishes.

In the first case we transpose k to the right-hand side and raise both sides of the equality to the second power. After some manipulations we obtain

$$2lm\sqrt{MN} = k^2 - l^2M - m^2N,$$

that is, $\sqrt{MN}$ is a rational, which is incorrect. Hence, the first case can not take place.

In the second case we see from equality (20) that $\sqrt{M}$ or $\sqrt{N}$ is a rational, in contradiction to the imposed condition. Hence, the second case is also rejected.

We thus have to recognize that $l = m = 0$. Equality (20) then shows that $k = 0$ as well. ■

* This is an example of familiar casual usage: "square-root radical" means not only the symbol $\sqrt{\ }$ that stands for the operation of extracting the square root, but also refers to any number of the type $\sqrt{2}$, $\sqrt{3}$ etc.

Lemma 2. *If $\sqrt{M}$ and $\sqrt{N}$ are nonsimilar radicals, then the equality*

$$k + l\sqrt{M} + m\sqrt{N} + n\sqrt{MN} = 0 \tag{21}$$

is possible only if $k = l = m = n = 0$.

In a shorter form

$$k + l\sqrt{M} + m\sqrt{N} + n\sqrt{MN} = 0 \Leftrightarrow k = l = m = n = 0.$$

▶ Assume that $l \neq 0$, $m \neq 0$, and $n \neq 0$. We transform equality (21) to the following form:

$$l\sqrt{M} + m\sqrt{N} = -k - n\sqrt{MN}.$$

Let us square both sides of this equality. Simple transformations then give

$$2(lm - kn)\sqrt{MN} = k^2 + n^2MN - l^2M - m^2N,$$

that is, $\sqrt{MN}$ is a rational number, which is incorrect. We thus have to reject the assumption $l \neq 0$, $m \neq 0$, $n \neq 0$, that is, we have to assume that *at least one* of the coefficients, l, m, n vanishes. But in this case equality (21) reduces to (20) so that by virtue of Lemma 1 all the remaining coefficients vanish. ■

Lemma 3. *If $p + q\sqrt{M}$ is a root of the equation*

$$a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n = 0$$

with integral coefficients, then $p - q\sqrt{M}$ is also a root of this equation.

▶ Given:

$$a_0(p + q\sqrt{M})^n + a_1(p + q\sqrt{M})^{n-1} + \ldots + a_{n-1}(p + q\sqrt{M}) + a_n = 0. \tag{22}$$

To be proved:

$$a_0(p - p\sqrt{M})^n + a_1(p - q\sqrt{M})^{n-1} + \ldots + a_{n-1}(p - q\sqrt{M}) + a_n = 0. \tag{23}$$

Let us remove the parentheses in (22). The terms $(q\sqrt{M})^\alpha$ obtained thereby are subsumed under two types:

(1) α is even (including $\alpha = 0$). All these terms are rational. We denote their sum by k.

(2) α is odd. All these terms are of the form $s\sqrt{M}$. We denote their sum by $l\sqrt{M}$.

Equality (22) thus transforms to

$$k + l\sqrt{M} = 0. \tag{24}$$

Let us make similar transformations over equality (23) which is obtained from (22) by substituting $-q\sqrt{M}$ for $q\sqrt{M}$. This substitution does not affect the terms containing $q\sqrt{M}$ to even powers, while the terms containing $q\sqrt{M}$ to odd powers only have their sign reversed. Equality (23) will thus become

$$k - l\sqrt{M} = 0. \tag{25}$$

Equality (24) can only hold if $k = l = 0$.

Indeed, if $l \neq 0$, equality (24) indicates that $\sqrt{M}$ is a rational. And if $l = 0$, then $k = 0$ as well.

But if $k = l = 0$, then (25) also holds. ■

Let us outline once again the idea of the proof. Equalities (24) and (25) are the equalities (22) and (23) appropriately transformed. Equality (24) yields $k = l = 0$, and $k = l = 0$ implies that (25) holds.

Lemma 4. *If $p + q\sqrt{M} + r\sqrt{N}$, where $\sqrt{M}$ and $\sqrt{N}$ are nonsimilar radicals, is a root of an equation with integral coefficients, then the numbers $p \pm q\sqrt{M} \pm r\sqrt{N}$, regardless of the combination of signs, are also the roots of this equation.*

To contract the notations we denote

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n.$$

Given:

$$\begin{aligned} P(p + q\sqrt{M} + r\sqrt{N}) \equiv & a_0(p + q\sqrt{M} \\ & + r\sqrt{N})^n + a_1(p + q\sqrt{M} + r\sqrt{N})^{n-1} + \ldots \\ & + a_{n-1}(p + q\sqrt{M} + r\sqrt{N}) + a_n = 0. \end{aligned} \tag{26}$$

To be proved:

$$P(p \pm q\sqrt{M} \pm r\sqrt{N}) = 0.$$

▶ Now we remove the parentheses in (26). All the terms obtained thereby will be of the form

$$Ap^{\alpha}(q\sqrt{M})^{\beta}(r\sqrt{N})^{\gamma},$$

where A are coefficients and α, β, γ are non-negative integral exponents. We subsume these terms under four types:

Type	β	γ	Kind of the term
1	Even	Even	Rational
2	Even	Odd	$t\sqrt{N}$
3	Odd	Even	$u\sqrt{M}$
4	Odd	Odd	$v\sqrt{MN}$

When parentheses in (26) are removed, we obtain

$$P(p+q\sqrt{M}+r\sqrt{N}) \equiv k+l\sqrt{M}+m\sqrt{N}+n\sqrt{MN}=0.$$

If we substitute $-q\sqrt{M}$ for $q\sqrt{M}$, it will not change the type-1 and type-2 terms, while type-3 and type-4 terms will only have their signs reversed. Therefore, if

$$P(p+q\sqrt{M}+r\sqrt{N})=k+l\sqrt{M}+m\sqrt{N}+n\sqrt{MN},$$

then

$$P(p-q\sqrt{M}+r\sqrt{N})=k-l\sqrt{M}+m\sqrt{N}-n\sqrt{MN}.$$

By going through similar steps with the other combinations of signs, we find: if

$$P(p+q\sqrt{M}+r\sqrt{N})=k+l\sqrt{M}+m\sqrt{N}+n\sqrt{MN},$$

then

$$P(p+q\sqrt{M}-r\sqrt{N})=k+l\sqrt{M}-m\sqrt{N}-n\sqrt{MN};$$
$$P(p-q\sqrt{M}+r\sqrt{N})=k-l\sqrt{M}+m\sqrt{N}-n\sqrt{MN};$$
$$P(p-q\sqrt{M}-r\sqrt{N})=k-l\sqrt{M}-m\sqrt{N}+n\sqrt{MN};$$

if $P(p+q\sqrt{M}+r\sqrt{N})=0$, then, by virtue of Lemma 2, $k=l=m=n=0$. But it implies that all the other values of $P(p\pm q\sqrt{M}\pm r\sqrt{N})$ vanish. ■

Let us have a look at examples.

Example 1. The number $1+\sqrt{2}$ is a quadratic irrational. How can an equation be found that generates this irrational.

Lemma 3 states that the number $1 - \sqrt{2}$ is also a root of this equation. Hence, the equation is

$$(x-1-\sqrt{2})(x-1+\sqrt{2})=0$$

or

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

Example 2. The number $\sqrt{2} + \sqrt{3}$ is not a quadratic irrational. The equation with integral coefficients that generates it has, by virtue of Lemma 4, the following roots:

$$x_1=\sqrt{2}+\sqrt{3};$$
$$x_2=\sqrt{2}-\sqrt{3};$$
$$x_3=-\sqrt{2}+\sqrt{3};$$
$$x_4=-\sqrt{2}-\sqrt{3}.$$

This equation is, therefore,

$$(x-\sqrt{2}-\sqrt{3})(x-\sqrt{2}+\sqrt{3})(x+\sqrt{2}-\sqrt{3})$$
$$\times(x+\sqrt{2}+\sqrt{3})=0,$$

or

$$x^4 - 10x^2 + 1 = 0.$$

Note 1. If the roots are known, the equation can obviously be found also by Vieta's formulas. For the normalized equation of fourth degree

$$x^4 + p_1x^3 + p_2x^2 + p_3x + p_4 = 0,$$

Vieta's formulas are

$$\left.\begin{aligned} p_1&=-(x_1+x_2+x_3+x_4);\\ p_2&=x_1x_2+x_1x_3+x_1x_4+x_2x_3+x_2x_4+x_3x_4;\\ p_3&=-(x_2x_3x_4+x_1x_3x_4+x_1x_2x_4+x_1x_2x_3);\\ p_4&=x_1x_2x_3x_4. \end{aligned}\right\}$$

Note 2. The reader may be somewhat surprised if he tries to check whether the equation indeed corresponds to the prescribed roots. In fact, the equation gives:

$$x=\pm\sqrt{5\pm2\sqrt{6}}.$$

At the first glance, this differs from the given roots $\pm\sqrt{2}\pm\sqrt{3}$. But in fact $\sqrt{3}\pm\sqrt{2}=\sqrt{5\pm2\sqrt{6}}$. This can be

checked either by squaring both sides of the equality or by employing the so-called formula for transforming complex radicals:

$$\sqrt{A \pm \sqrt{B}} = \sqrt{\frac{A+\sqrt{A^2-B}}{2}} \pm \sqrt{\frac{A-\sqrt{A^2-B}}{2}}. \quad (27)$$

Formula (27) helps only when $A^2 - B$ is an exact square. This is not the case in our example.

4.3.3. Euler's Theorem. A nonterminating continued fraction is said to be *periodical* if its terms form a periodical sequence. For example, such are the fractions

$$[0;\ 1,\ 1,\ 1,\ \ldots];$$

$$[2;\ 1,\ 5,\ 1,\ 5,\ 1,\ 5,\ \ldots];$$

$$[0;\ 1,\ 2,\ 3,\ 5,\ 3,\ 5,\ 3,\ 5,\ \ldots].$$

The first two fractions are *purely periodical*, and the third is a *mixed periodical* continued fraction. In this classification we ignore the integral component a_0. What follows is a more straightforward definition.

A nonterminating continued fraction is said to be periodical if there exist natural numbers N and k such that

$$a_{n+k} = a_n$$

for any $n \geqslant N$.

The following theorem, proved by Leonard Euler in 1737, holds for continued fractions.

Theorem. *The value of each periodical continued fraction is a quadratic irrational.*

▶ Let us consider two examples.

Example 1. $[0;\ 1,\ 1,\ 1,\ \ldots]$. We have

$$\alpha = \cfrac{1}{1+\cfrac{1}{1+\ddots}}$$

Let us apply to this equality an operation of "rewinding", consisting of (1) taking the reciprocal of each side, and (2) subtracting the integral part (entier) from each side. In this

particular case these steps are only made once:

$$\frac{1}{\alpha} = 1 + \cfrac{1}{1+\cfrac{1}{1+\ddots}};$$

$$\frac{1}{\alpha} - 1 = \cfrac{1}{1+\cfrac{1}{1+\ddots}}.$$

What we have now on the right is the initial fraction, that is, α:

$$\frac{1}{\alpha} - 1 = \alpha,$$

which gives us a quadratic equation for α:

$$\alpha^2 + \alpha - 1 = 0,$$

whence $\alpha = -\frac{1}{2} + \frac{1}{2}\sqrt{5}$ (of course, the negative root must be rejected).

This analysis shows that any fraction of the type [0; a, a, a, . . .] represents a quadratic irrational.

And what if the period consists not of one numerical but of k numerals? Then k pairs of steps will be made in "rewinding".

Example 2. [0; 1, 2, 1, 2, . . .].

$$\alpha = \cfrac{1}{1+\cfrac{1}{2+\cfrac{1}{1+\cfrac{1}{2+}}}}$$

$$\frac{1}{\alpha} - 1 = \cfrac{1}{2+\cfrac{1}{1+\cfrac{1}{2+\ddots}}},$$

$$\frac{1}{\frac{1}{\alpha}-1} - 2 = \cfrac{1}{1+\cfrac{1}{2+\ddots}}$$

$$\frac{1}{\frac{1}{\alpha}-1} - 2 = \alpha$$

or

$$\alpha^2+2\alpha-2=0$$

whence

$$\alpha=-1+\sqrt{3}.$$

Note that the root cannot happen to be rational whatever the period because the initial continued fraction is nonterminating.

But what happens if a_0 is nonzero? In this case we transpose a_0 to the left-hand side, and begin "rewinding". ■

However, this method is cumbersome if the period is long. For this reason we shall give another proof, not as lucid as the one above but a short one.

▶ Let a nonterminating continued fraction $\alpha = [0; a_1, a_2, \ldots]$ be a purely periodical one, with the period length equal to k. Then $\alpha = \alpha_{k+1}$ (recall that α_{k+1} is the $(k+1)$th complete quotient):

$$\alpha=[0;\ a_1,\ a_2,\ \ldots,\ a_k,\ \underbrace{a_1,\ a_2,\ \ldots}_{\alpha_{k+1}}].$$

Formula (9) yields

$$\alpha=\frac{p_k\alpha_{k+1}+p_{k-1}}{q_k\alpha_{k+1}+q_{k-1}}.$$

Hence,

$$\alpha=\frac{p_k\alpha+p_{k-1}}{q_k\alpha+q_{k-1}},$$

that is, α satisfies a quadratic equation

$$q_kx^2+(q_{k-1}-p_k)\,x-p_{k-1}=0. \tag{28}$$

The roots of this equation have opposite signs, and α is the positive root.

If the fraction is a mixed periodical one,

$$\alpha=[a_0;\ a_1,\ a_2,\ \ldots,\ a_N,\ \underbrace{a_{N+1},\ \ldots,\ a_{N+k}}_{\text{period}},\ \ldots],$$

we must first "rewind" from right to left the first part of the fraction up to the term a_N, inclusive, and then apply the proof as given above. ■

Note. The number α is irrational because it is represented by a nonterminating continued fraction. Consequently, the discriminant of equation (28) must not be an exact square. This statement can be tested by a direct evaluation:

$$D=(p_k-q_{k-1})^2+4p_{k-1}q_k=p_k^2-2p_kq_{k-1}+q_{k-1}^2+4p_{k-1}q_k=\ldots$$

By adding and subtracting the term $4p_kq_{k-1}$ we find:

$$\ldots=p_k^2+2p_kq_{k-1}+q_{k-1}^2-4p_kq_{k-1}+4p_{k-1}q_k$$

$$=(p_k+q_{k-1})^2-4q_{k-1}q_k\left(\frac{p_k}{q_k}-\frac{p_{k-1}}{q_{k-1}}\right)=\ldots$$

At this juncture we apply formula (13):

$$\ldots = (p_k + q_{k-1})^2 + 4 \cdot (-1)^k.$$

Finally, we have

$$D = (p_k + q_{k-1})^2 + 4 \cdot (-1)^k$$

or

$$D - (p_k + q_{k-1})^2 = \pm 4.$$

We see that D is not an exact square. The difference between squares of natural numbers cannot equal 4. If the set of natural numbers is supplemented with zero, there will be found a *unique* pair of squares spaced by 4: 0 and 4.

4.3.4. Lagrange Theorem. As you could see in the preceding subsection, Euler's theorem is proved quite easily. The inverse theorem is considerably more difficult to prove. It has been proved by Lagrange in 1770.

Lagrange Theorem. *Each quadratic irrational is represented by a periodic continued fraction.*

Lagrange succeeded in proving the theorem in a very complicated manner. Quite a few mathematicians attempted to simplify the proof but to return the original Lagrange's idea. A hundred-odd years later the French mathematician Charves suggested a simpler proof based on a different idea. First we shall outline Charves' idea, and then give the detailed proof.

Let α be a quadratic irrational. Let us expand it into a continued fraction interrupting the process at each step, beginning with the second step:

$$\alpha = [a_0;\ a_1 \mid \alpha_2] = [a_0;\ a_1,\ a_2 \mid \alpha_3]$$
$$= \ldots = [a_0;\ a_1,\ a_2,\ \ldots,\ a_{n-1} \mid \alpha_n] = \ldots$$

Here $\alpha_2, \alpha_3, \ldots, \alpha_n, \ldots$ are complete quotients. We were able to see in Subsection 3.1.5 that if some complete quotient happens to repeat, that is, if we find that $\alpha_n = \alpha_{n+k}$, the continued fraction will be periodical.

We shall prove, first, that each term satisfies a quadratic equation with integral coefficients:

$$A_n \alpha_n^2 + B_n \alpha_n + C_n = 0. \tag{29}$$

Of course, equation (29) can vary for different values of α_n, and for this reason the coefficients A, B, C are equipped with subscripts. Rather, we should say: each α_n satisfies *its own* quadratic equation with integral coefficients.

Second, we shall prove that the magnitudes of the coefficients in (29) are bounded:*

$$\left.\begin{aligned} |A_n| &< L; \\ |B_n| &< M \\ |C_n| &< N. \end{aligned}\right\} \tag{30}$$

* We assume that a quadratic equation with integral coefficients is written in the form irreducible to lower terms. Otherwise this statement would be meaningless.

Attention, please! Here is Charves' clever idea. *The bounds L, M, N are independent of n* (they depend exclusively on α). Since A_n, B_n, C_n are integers, only a *finite* number of admissible values exist for them. Consequently, for each given α the number of possible equations (29) and hence, the number of possible roots of these equations is finite. Obviously, the sequence of complete quotients $\alpha_2, \alpha_3, \ldots, \alpha_n, \ldots$ will inevitably start to repeat itself; this has to be proved.

▶ Now let us implement this plan. First we prove (29), and then (30).

The quadratic irrational α satisfies a certain quadratic equation with integral coefficients.

$$A\alpha^2 + B\alpha + C = 0. \tag{31}$$

By virtue of (9),

$$\alpha = \frac{p_{n-1}\alpha_n + p_{n-2}}{q_{n-1}\alpha_n + q_{n-2}}. \tag{32}$$

We substitute (32) into (31) and eliminate the denominator:

$$A(p_{n-1}\alpha_n + p_{n-2})^2 + B(p_{n-1}\alpha_n + p_{n-2})(q_{n-1}\alpha_n + q_{n-2}) + C(q_{n-1}\alpha_n + q_{n-2})^2 = 0$$

or

$$A_n\alpha_n^2 + B_n\alpha_n + C_n = 0,$$

where

$$\left.\begin{aligned} A_n &= Ap_{n-1}^2 + Bp_{n-1}q_{n-1} + Cq_{n-1}^2; \\ B_n &= 2Ap_{n-1}p_{n-2} + B(p_{n-1}q_{n-2} + p_{n-2}q_{n-1}) + 2Cq_{n-1}q_{n-2}; \\ C_n &= Ap_{n-2}^2 + Bp_{n-2}q_{n-2} + Cq_{n-2}^2. \end{aligned}\right\} \tag{33}$$

It remains for us to prove that coefficients (33) have bounded magnitudes. From (14),

$$\left|\frac{p_{n-1}}{q_{n-1}} - \alpha\right| < \frac{1}{q_{n-1}^2}.$$

This can be rewritten in the form

$$\frac{p_{n-1}}{q_{n-1}} - \alpha = \frac{\delta}{q_{n-1}^2},$$

where $-1 < \delta < 1$, whence

$$p_{n-1} = \alpha q_{n-1} + \frac{\delta}{q_{n-1}} \quad (-1 < \delta < 1).$$

Let us substitute this expression for p_{n-1} into the first formula of (33):

$$A_n = A\left(\alpha q_{n-1} + \frac{\delta}{q_{n-1}}\right)^2 + B\left(\alpha q_{n-1} + \frac{\delta}{q_{n-1}}\right) q_{n-1} + Cq_{n-1}^2$$

$$= q_{n-1}^2(A\alpha^2 + B\alpha + C) + 2A\delta + B\delta + \frac{A\delta^2}{q_{n-1}^2} = \ldots$$

Obviously, the expression in parentheses vanishes owing to (31):

$$\ldots = \left(2A\alpha + B + \frac{A\delta}{q_{n-1}^2}\right)\delta.$$

But $|\delta| < 1$, so that

$$|A_n| < \left|2A\alpha + B + \frac{A\delta}{q_{n-1}^2}\right|.$$

We shall also take into account that $q_{n-1}^2 > 1$ ($q_0 = 1$, and the sequence q_n is strictly increasing). The inequality is only strengthened if we drop q_{n-1}^2 (i.e. replace it by unity):

$$|A_n| < |2A\alpha + B + A\delta| \leqslant |2A\delta| + |B| + |A| \cdot |\delta|$$
$$< |2A\alpha| + |B| + |A|.$$

Our goal is attained: we have indicated for $|A_n|$ a bound independent of n.

Instead of conducting similar manipulations for $|C_n|$, we note that C_n is obtained from A_n by replacing n by $n - 1$, that is, $C_n = A_{n-1}$. The bound established above is independent of n, and thus is valid for C_n. As for B_n, a detour will be more effective. Let us calculate the discriminant of equation (29) by using formula (33). We omit here a long but dull series of manipulations* and give the final result:

$$B_n^2 - 4A_nC_n = (p_{n-1}q_{n-2} - p_{n-2}q_{n-1})^2)(B^2 - 4AC).$$

But formula (12) states that

$$p_{n-1}q_{n-2} - p_{n-2}q_{n-1} = (-1)^{n-2}.$$

Hence,

$$B_n^2 - 4A_nC_n = B^2 - 4AC. \tag{34}$$

This formula expresses a natural fact: when a quadratic irrational α is expanded into a continued fraction, the complete quotients are quadratic irrationals of the same nature as α is. This nature is determined by the discriminant. They are all of the type

$$\alpha_n = s_n + t_n\sqrt{D}$$

for the same D.

Now we conclude from (34) that

$$B_n^2 = B^2 - 4AC + 4A_nC_n.$$

All terms in the right-hand side being bounded, B_n^2 and with it $|B_n|$ are bounded as well. ■

Euler's and Lagrange theorems can be merged in the following formulation: *Quadratic irrationals, and only they, are represented by periodical continued fractions.*

---

* We leave these manipulations to the reader. A mathematician should be patient and unafraid of long chains of transformations.

Chapter 5

Approximation of Real Numbers

5.1. Approximation by Convergents

5.1.1. High-Quality Approximation. The tiring march has finally brought us to the goal of our journey. This chapter will disclose what purpose is served by continued fractions.

In Section 1.1.1 we have interpreted the problem of approximation in a very broad sense. Now we switch to a more specific problem. Take the set **R** of real numbers*, and single out in **R** the subset M_q of all fractions with denominators not greater than q. The problem is to find for each number $\alpha \in \mathbf{R}$ the closest to it number $r \in M_q$.

Assume now that we were able to find such a number, that is, we found an approximation $\alpha \approx r$. The utility of this approximation consists in that *the accuracy cannot be improved without increasing the denominator*: indeed, r is the closest to α number in M_q.

Note that if we chose to take the set of fractions with denominators *exactly equal to* q, this approximation would not, in general, have high quality in the sense outlined above. For example, we see from Table 1 of Subsection 1.1.4 that the approximation of π in units of 1/10 has low quality in comparison with larger fractional units, namely, 1/9, 1/8, 1/7, and 1/6.

The concept of "quality" does not have a single sharply defined meaning in approximation theory, so that we have to specify each time in what sense this term is being used.

5.1.2. The Main Property of Convergents. We can define the best rational approximation of a number α as a fraction $\frac{p}{q}$ which provides a lower absolute error than any other

---

* It is sufficient to consider only the set of positive real numbers because nothing principally new arises from adding negative numbers: if $\pi \approx \frac{22}{7}$, then $-\pi \approx -\frac{22}{7}$.

fraction with a denominator $\leqslant q$ (obviously, in this sense, the best approximation is not unique). With this definition, *convergents provide the best approximations for a continued fraction.* Sometimes this property is said to constitute the *main property of convergents.* Let us give it the following formulation:

Theorem. *If* $\frac{p_n}{q_n}$ $(n \geqslant 1)$ *is a convergent for a number* α, *an* $\frac{p}{q}$ *is any other fraction with* $q \leqslant q_n$, *then*

$$\left|\alpha - \frac{p_n}{q_n}\right| < \left|\alpha - \frac{p}{q}\right|.$$

This means that the convergent gives an approximation that cannot be improved without increasing the denominator.

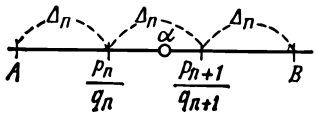

Fig. 12

▶ Consider two cases: (1) $q < q_n$ and (2) $q = q_n$ (we shall see that the second case is trivial).

(1) The number α belongs to a segment between two convergents $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$ (Fig. 12). The length of this segment is $|\Delta_n| = \frac{1}{q_n q_{n+1}}$. The point α may either be an internal point of this segment or coincide with $\frac{p_{n+1}}{q_{n+1}}$ (if α is a rational and $\frac{p_{n+1}}{q_{n+1}}$ is the last convergent). Therefore,

$$\left|\alpha - \frac{p_n}{q_n}\right| \leqslant |\Delta_n|.$$

Let $\frac{p}{q}$ be an arbitrary fraction whose denominator is less than q_n, and hence, certainly less than q_{n+1}:

$$q < q_n < q_{n+1}.$$

Let us evaluate the distance from $\frac{p}{q}$ to the ends of the segment $\left[\frac{p_n}{q_n}, \frac{p_{n+1}}{q_{n+1}}\right]$:

$$\left|\frac{p}{q}-\frac{p_n}{q_n}\right|=\frac{|pq_n-p_nq|}{qq_n}\geqslant\frac{1}{qq_n};$$
$$\left|\frac{p}{q}-\frac{p_{n+1}}{q_{n+1}}\right|=\frac{|pq_{n+1}-p_{n+1}q|}{qq_{n+1}}\geqslant\frac{1}{qq_{n+1}}.$$

These inequalities are only strengthened if q in the right-hand sides are replaced by q_n and q_{n+1}:

$$\left.\begin{aligned}\left|\frac{p}{q}-\frac{p_n}{q_n}\right|&>\frac{1}{q_{n+1}q_n}=|\Delta_n|;\\ \left|\frac{p}{q}-\frac{p_{n+1}}{q_{n+1}}\right|&>\frac{1}{q_nq_{n+1}}=|\Delta_n|.\end{aligned}\right\}\qquad(35)$$

Inequalities (35) signify that each end of segment $\left[\frac{p_n}{q_n}, \frac{p_{n+1}}{q_{n+1}}\right]$ and point $\frac{p}{q}$ are separated by a distance greater than the length of this segment, $|\Delta_n|$. Marking off the segment $|\Delta_n|$ to the left and to the right of points $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$ (Fig. 12) we obtain the forbidden zone $[AB]=\left[\frac{p_n}{q_n}-\Delta_n, \frac{p_{n+1}}{q_{n+1}}+\Delta_n\right]$ in which the fraction $\frac{p}{q}$ cannot fall (since points A and B are also forbidden). Now it is clear that $\frac{p}{q}$ is a poorer approximation for α than $\frac{p_n}{q_n}$. Indeed,

$$\left|\alpha-\frac{p_n}{q_n}\right|<|\Delta_n|;$$
$$\left|\alpha-\frac{p}{q}\right|>|\Delta_n|.$$

Hence

$$\left|\alpha-\frac{p_n}{q_n}\right|<\left|\alpha-\frac{p}{q}\right|\qquad(q<q_n).$$

(2) Now we shall analyse the case of $q=q_n$. Is it possible for another fraction with the same denominator to provide a better or an equally good approximation than the convergent? In other words, can it happen that

$$\left|\alpha-\frac{p}{q_n}\right|\leqslant\left|\alpha-\frac{p_n}{q_n}\right|\qquad(p\neq p_n).$$

For the sake of definiteness we assume that $\frac{p_n}{q_n}$ lies to the left of α (Fig. 13), that is, n is even (arguments are quite similar if n is odd). Can α be closer to $\frac{p_n+1}{q_n}$ than to $\frac{p_n}{q_n}$, or at

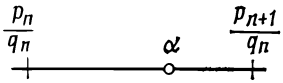

Fig. 13

least lie in the middle, that is, is it possible that

$$\frac{p_n+1}{q_n}-\alpha\leqslant\alpha-\frac{p_n}{q_n}\ ? \tag{36}$$

This is equivalent to

$$\alpha-\frac{p_n}{q_n}\geqslant\frac{1}{2q_n}. \tag{37}$$

On the other hand, we know that

$$\alpha-\frac{p_n}{q_n}\leqslant\frac{1}{q_nq_{n+1}}.$$

As follows from inequalities (36) and (37),

$$\frac{1}{2q_n}\leqslant\frac{1}{q_nq_{n+1}},$$

that is, $q_{n+1}\leqslant 2$.

It means that inequality (36) is only possible if $q_{n+1}=1$ or $q_{n+1}=2$. This situation occurs only for $n=0$ and $n=1$.

The following example shows that inequality (36) can indeed hold under these conditions:

$$[2;\,2]=2+\frac{1}{2}.$$

In this example $\frac{p_0}{q_0}=\frac{2}{1}$. The fraction $\frac{3}{1}$ although it is not a convergent provides an approximation that is just as good. No such example can be given for $n=1$ because the last term of a continued fraction cannot be a unity. ■

Note that the converse theorem does not hold, that is, the property proved above is not exclusive to convergents. There exist fractions that are not convergents but nevertheless give a better approximation of a number α than any fraction with a smaller denominator. For example, in Subsection 3.1.1 we

have listed the convergents of $\frac{61}{27} = [2; 3, 1, 6]$. The reader can verify* that the approximations

$$\frac{61}{27} \approx \frac{34}{15} \text{ and } \frac{61}{27} \approx \frac{43}{19}$$

are the best. The absolute errors for them,

$$\left|\frac{61}{27} - \frac{34}{15}\right| \approx 0.007, \quad \left|\frac{61}{27} - \frac{43}{19}\right| \approx 0.004,$$

are less than those for any fraction with a smaller denominator, although these two fractions are not convergents. The theorem as proved does not mean, therefore, that convergents provide the best approximation of real numbers.

5.1.3. Convergents Have the Highest Quality. If the quality of approximation is evaluated in a different sense, as described in Subsection 1.1.4 then convergents have no competitors.

The point is that it would be unfair to evaluate the quality of an approximation by the quantity $\left|\alpha - \frac{p}{q}\right|$ independently of the magnitude of the fractional unit. Higher accuracy can be demanded from smaller fractional units (i.e. from larger q). Consequently, it is desirable to estimate the absolute error $\left|\alpha - \frac{p}{q}\right|$ on a q-dependent scale, for example, multiplying $\left|\alpha - \frac{p}{q}\right|$ by q. The result is the normalized absolute error [see formula (1)]

$$h = |q\alpha - p|$$

or the quality factor [see formula (2)]

$$\lambda = \frac{1}{2h} = \frac{1}{2|q\alpha - p|}.$$

Judging by these characteristics it is the convergents, and *nothing but them*, that have the highest quality: the normalized absolute error of a convergent is smaller (and hence, the quality factor is greater) than in all other fractions with smaller (or identical) denominators.

But this is not yet the end. Convergents are found to have a quality higher than not only that of fractions with smaller or equal denominators but even of fractions with the nearest greater denominators: quality will not be increased by in-

* See footnote on p. 71.

creasing the denominator until we come to the denominator of the next convergent.

These arguments hide two trivial exceptions that will surface in the course of analysis.

Now we shall summarize the above reasoning into two theorems converse to each other.

Theorem 1. *If* $\frac{p_n}{q_n}$ *is a convergent for a number* α *and* $\frac{p}{q}$ *is an arbitrary fraction with* $q < q_{n+1}$, *then*

$$| q_n\alpha - p_n | \leqslant | q\alpha - p |.$$

The equality sign occurs only if: (1) $\alpha = \frac{p_{n+1}}{q_{n+1}}$, *that is,* $\frac{p_n}{q_n}$ *is the last-but-one convergent, and* (2) $n = 0$, $\alpha = [a_0; 2]$.

▶ Note that $\frac{p}{q}$ is a *different* fraction, that is, we eliminate the case of no interest, $\frac{p}{q} = \frac{p_n}{q_n}$. Hereafter we assume that the fraction $\frac{p}{q}$ is irreducible to lower terms.

We shall consider separately two cases: (1) $0 < q < q_{n+1}$, $q \neq q_n$, and (2) $q = q_n$.

(1) Let us represent p and q by identical linear combinations (with identical coefficients) of the corresponding terms of the convergents $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$, that is,

$$\left.\begin{aligned} q_n x + q_{n+1} y &= q; \\ p_n x + p_{n+1} y &= p. \end{aligned}\right\} \tag{38}$$

This system yields the coefficients x and y.

The determinant of (38), $p_{n+1}q_n - p_n q_{n+1}$, is recognizable from formula (12):

$$D_n = p_{n+1}q_n - p_n q_{n+1} = (-1)^n.$$

System (38) determines the pair of numbers x and y in a single-valued manner because $D_n \neq 0$. Besides, $| D_n | = 1$, and we conclude that x and y are integers.

Both x and y are nonzero. Indeed, if $x = 0$, then system (38) gives $y = 1$ (because both fractions $\frac{p}{q}$ and $\frac{p_{n+1}}{q_{n+1}}$ are irreducible to lower terms) and $q = q_{n+1}$, in contradiction with the imposed condition. And if $y = 0$, we likewise obtain $q = q_n$, which is the case to be analysed later.

The coefficients x and y cannot be of like signs. If $x > 0$ and $y > 0$, then the first equation in (38) would give $q > q_{n+1}$. If $x < 0$ and $y < 0$, then p and q would be negative. Hence, x and y are of unlike signs.

In order to find the normalized absolute error for the fraction $\frac{p}{q}$, we begin with multiplying the first equation by α and subtracting

from it the second equation:

$$(q_n\alpha - p_n)\, x + (q_{n+1}\alpha - p_{n+1})\, y = q\alpha - p. \tag{39}$$

The differences in the parentheses in the left-hand side of equation (39) are of unlike signs (because the convergents $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$ aproximate α on the opposite sides). The numbers x and y are also of unlike signs. Both terms in the left-hand side are therefore positive (more precisely, the first is strictly positive and the second is nonnegative). Hence,

$$| q_n\alpha - p_n | \cdot | x | + | q_{n+1}\alpha - p_{n+1} | \cdot | y | = | q\alpha - p |.$$

Therefore,

$$| q_n\alpha - p_n | \leqslant | q\alpha - p |, \tag{40}$$

which as to be proved.

Now we shall determine the conditions corresponding to the equality sign in (40). As follows from the analysis above, it is possible only if

$$\left.\begin{aligned} q_{n+1}\alpha - p_{n+1} &= 0; \\ | x | &= 1. \end{aligned}\right\} \tag{41}$$

Let us analyse the case (41) in more detail. If $x = 1$, we shall have $y < 0$. But then the first equation in (38) would yield $q < 0$. Hence, $x \neq 1$ and therefore, $x = -1$. This entails $y = 1$. Indeed, if we assume $y > 1$, then the first equation in (38) can be rewritten as follows:

$$-q_n + q_{n+1} + q_{n+1}(y-1) = q$$

which implies $q > q_{n+1}$.

The mandatory situation in the case (41) is thus $x = -1$, $y = 1$, that is,

$$\left.\begin{aligned} q &= q_{n+1} - q_n; \\ p &= p_{n+1} - p_n. \end{aligned}\right\} \tag{42}$$

This entails the equality sign in (40).

Note that the first condition in (42) can be transformed to

$$q = a_{n+1}q_n + q_{n+1} - q_n = (a_{n+1} - 1)\, q_n + q_{n-1}.$$

In the case under consideration a_{n+1} is the last term of the continued fraction, and hence, $a_{n+1} \geqslant 2$. The last equality therefore implies

$$q > q_n.$$

Consequently, the equality in (40) cannot occur when $q < q_n$.

(2) Now we shall consider the case of $q = q_n$. We already know from Subsection 5.1.1 that in this case, for $p \neq p_n$,

$$\left| \alpha - \frac{p_n}{q_n} \right| < \left| \alpha - \frac{p}{q_n} \right|.$$

By multiplying both sides of this inequality by q_n we obtain

$$| q_n\alpha - p_n | < | q_n\alpha - p |,$$

which completes the prove (of course, the exceptional case, possible

when $n = 0$, is retained here as well). The theorem has thus been proved for all $q < q_{n+1}$. ■

Theorem 2 (converse). *If the normalized absolute error for a number α and a fraction $\frac{p}{q}$ is less than that for any other fraction $\frac{p'}{q'}$ and $q' \leqslant q$, then $\frac{p}{q}$ is a convergent for α.*

▶ We assume, as always, that the fraction $\frac{p}{q}$ is irreducible to lower terms. Besides, if α is a rational, $\alpha = \frac{p_{n+1}}{q_{n+1}}$, then q cannot be greater than q_{n+1} because the normalized absolute error for the fraction $\frac{p_{n+1}}{q_{n+1}}$ is zero, while according to the initial condition it should be greater than $|q\alpha - p|$.

Assume that $\frac{p}{q}$ is not a convergent. Then its denominator lies somewhere between the denominators of two neighbouring convergents, that is,

$$q_n < q < q_{n+1}.$$

The direct theorem then gives

$$|q_n\alpha - p_n| < |q\alpha - p|.$$

This contradicts the condition stated in the theorem: since $q_n < q$, then $\frac{p}{q}$ must give a smaller normalized absolute error than $\frac{p_n}{q_n}$. The assumption that $\frac{p}{q}$ is not a convergent is therefore false. ■

Note 1. We have proved that *convergents, and only convergents, provide a smaller normalized absolute error and hence, a greater quality factor, than all other fractions with smaller denominators.*

Why "with smaller denominators" only? Could it be true for fractions with slightly greater denominators?

No, it could not. Only the direct theorem holds for denominators q in the interval $q_n < q < q_{n+1}$, and it cannot be converted.

Note 2. Let us consider in more detail the case (41):

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

We shall directly demonstrate that although the fraction

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

is not a convergent, and although $q_n < q < q_{n+1}$, this fraction nevertheless has the same quality as the convergent $\frac{p_n}{q_n}$.

The manipulations that follow need not be explained:

$$| q\alpha - p | = \left| (q_{n+1} - q_n) \frac{p_{n+1}}{q_{n+1}} - p_{n+1} + p_n \right|$$

$$= \left| \frac{p_n q_{n+1} - q_n p_{n+1}}{q_{n+1}} \right| = \frac{1}{q_{n+1}};$$

$$| q_n\alpha - p_n | = \left| q_n \frac{p_{n+1}}{q_{n+1}} - p_n \right| = \left| \frac{q_n p_{n+1} - p_n q_{n+1}}{q_{n+1}} \right| = \frac{1}{q_{n+1}},$$

that is, $| q\alpha - p | = | q_n\alpha - p_n |$. Recall that this would be impossible in the case $q < q_n$: it implies $| q\alpha - p | < | q_n\alpha - p_n |$.

For example, the consecutive convergents for the fraction $\alpha = \frac{61}{27}$ are (see Subsection 3.1.1)

$$\frac{p_0}{q_0} = \frac{2}{1}, \quad \frac{p_1}{q_1} = \frac{7}{3}, \quad \frac{p_2}{q_2} = \frac{9}{4}, \quad \frac{p_3}{q_3} = \frac{61}{27}.$$

The fraction $\frac{p}{q} = \frac{p_3 - p_2}{q_3 - q_2} = \frac{52}{23}$ has the same quality as $\frac{9}{4}$ despite the fact that $4 < 23 < 27$.

Note 3. Let us compare the approximations of the fraction $\alpha = \frac{61}{27}$ by fractions with denominators 1, 2, 3, 4 (Table 2).

Table 2

q	Approximate value of α	Normalized absolute error h	Quality factor γ
1	$\frac{2}{1}$	$\frac{7}{27}$	$\frac{27}{14} = 1\frac{13}{14}$
2	$\frac{5}{2}$	$\frac{13}{27}$	$\frac{27}{26} = 1\frac{1}{26}$
3	$\frac{7}{3}$	$\frac{2}{9}$	$\frac{9}{4} = 2\frac{1}{4}$
4	$\frac{9}{4}$	$\frac{1}{27}$	$\frac{27}{2} = 13\frac{1}{2}$

By looking at this table we can determine, *without expanding the number* $\frac{61}{27}$ *into a continued fraction*, that $\frac{9}{4}$ is its convergent: the quality factor of $\frac{9}{4}$ is greater than all preceding factors. The same is true for the fraction $\frac{7}{3}$. However, $\frac{5}{2}$ is not a convergent because its quality factor is less than that of $\frac{2}{1}$.

Chapter 6

Solutions

6.1. The Mystery of Archimedes' Number

6.1.1. The Key to All Puzzles. The readers who have worked their way through Chapters 1, 2, 3, 4 and 5 will now be rewarded. We are ready to explain the puzzles of Chapter 1.

In fact, this booklet has been written for the sake of one short conclusion: *If you want to approximate a real number with high accuracy by a fraction with sufficiently simple denominator, replace it with convergents.*

Thus we solve both Archimedes' problem and the problem of the calendar.

Note that Christian Huygens came to continued fractions when trying to approximate real numbers by sufficiently simple fractions. He needed to construct a model of the Solar system in which planets were modelled by gear-wheels. To reproduce revolution periods with sufficiently high accuracy wheels had to have staggeringly high numbers of teeth. Huygens looked for, and found, a general method of solving such problems: the substitution of much smaller numbers for the large ones, reproducing their ratios as accurately as possible. In this way he invented continued fractions as an auxiliary tool, and discovered many of their properties although Raffael Bombelli in Italy had operated with these fractions (in a more superficial way) a hundred years earlier.

N. N. Luzin used to say in such cases that "even chips and shavings are valuable in the laboratory of a great scientist."

6.1.2. The Secret of Archimedes' Number. To find approximations of the number π we expand it into a continued fraction. We can take its decimal approximation with high accuracy, for example, $3.14159265 = \frac{314159265}{100000000}$, and apply Euclid's algorithm:

$$\pi = [3;\ 7,\ 15,\ 1,\ 288,\ 1,\ \ldots].$$

Now we calculate convergents by the method of Subsection 3.1.4:

n	0	1	2	3	4
a_n	3	7	15	1	288
p_n	3	22	333	355	102 595
q_n	1	7	106	113	32 657

And this is all, as simple as that. This table exposes Archimedes' secret, as well as that of Metius. The table demonstrates:

Approximation	Convergent $\frac{p_n}{q_n}$
0th	3 (by defect)
1st	$\frac{22}{7}$ (by excess)
2nd	$\frac{333}{106}$ (by defect)
3rd	$\frac{355}{113}$ (by excess)

Can it really be said that Archimedes and Metius are at last "exposed": they had employed continued fractions, with Archimedes using the convergent $\frac{p_1}{q_1}$, and Metius the convergent $\frac{p_3}{q_3}$?

No, it cannot, at least not about Archimedes.

It should be clear to the reader that the problem we have solved is one of mathematics, not of history. We have demonstrated how one *could* have come to the approximation of π by a fraction $\frac{22}{7}$, but this does not mean that Archimedes did use this approach. In fact, it cannot be ruled out that he had used the continued fractions algorithm. This conjecture is

supported by two arguments: (1) this is the most natural approach when decimals have not yet been invented and (2) the ancients preferred fractions with unity for the numerator. Only such fractions were in use in Egypt and Babylon, and other fractions were gaining recognition only very slowly. Nevertheless, these are merely speculative arguments, and would be rejected at any court of law. No direct evidence has been found. In order to evaluate π Archimedes calculated the perimeters of inscribed and circumscribed regular polygons, using the "duplication formula". We do not know how Archimedes had extracted roots, for he only gave the final result. Historians were unable to come to a universally acceptable conclusion on this subject.

The advantages of fractions with denominator 7 can be discovered empirically, while they are compared to fractions with different denominators. But Metius (or rather, Antoniszoon) could not act this way. It would hardly be possible to find the complicated fraction $\frac{355}{113}$ without a theory. There is virtually no doubt that Antoniszoon resorted to continued fractions. It is perfectly clear why he stopped with the convergent $\frac{355}{113}$. In fact, this is the last acceptable fraction. The next one, $\frac{102\,595}{32\,657}$, is so cumbersome that it cannot have any practical significance.

6.2. The Solution to the Calendar Problem

6.2.1. The Use of Continued Fractions. Let us think first how we ourselves would solve the problem of alternation of ordinary and leap years. We would represent the duration of the year by a continued fraction

$$\begin{aligned} 1 \text{ year} &= 365 \text{ days } 5 \text{ hours } 48 \text{ minutes } 46 \text{ seconds} \\ &= [365;\ 4,\ 7,\ 1,\ 3,\ 5,\ 64) \text{ days.} \end{aligned}$$

Note 1. The number π is an irrational and is represented by a nonterminating continued fraction. The length of the year is an empirical quantity. All empirical quantities are measured with errors, so it would be meaningless to consider them as rational or irrational. The length of the year as given above is the *adopted value* and we have to treat it as exact. It is given by a terminating continued fraction.

Note 2. We need not express the length of the year by a decimal in fractions of one day (by analogy with what we did in the case of π) if we wish to represent this duration by a continued fraction. The calculations are carried out as follows (we have dropped the integral component):

$$\frac{5\text{ h }48\text{ m }46\text{ s}}{1\text{ day}}=\frac{20\,926\text{ s}}{86\,400\text{ s}}=\frac{10\,463}{43\,200};$$

$$43\,200=4\cdot 10\,463+1348;$$

$$10\,463=7\cdot 1348+1027;$$

$$1348=1\cdot 1027+321;$$

$$1027=3\cdot 321+64;$$

$$321=5\cdot 64+1;$$

$$64=64\cdot 1.$$

Let us find several convergents of the continued fraction representing the length of the year. The integral part can be omitted because we need to remind that each year contains 365 full days:

4	7	1	3	5
1	7	8	31	163
4	29	33	128	673

Each column gives a solution to the calendar problem. For example, the first column gives to the year an approximate length of $365\frac{1}{4}$ days. This duration is achieved by setting one year in four as a leap year. In general, the third row gives the length of the cycle (or period), and the second row gives the number of leap years per cycle. For example, the second column prescribes the following solution: seven leap years in a 29-years cycle. This corresponds to the average duration of the year of $365\frac{7}{29}$ days. This is a more accurate pattern than $365\frac{1}{4}$ but a more complicated one.

6.2.2. How to Choose a Calendar. Now it is clear that we are offered only four options.

In order to avoid misunderstandings, we have to remark that a very large number of calendars exist in the world.

There exist solar and lunar calendars. Different peoples use different starting points for counting off the years, different numbers of months per year (12 or 13), different (and tremendously diverse) starting dates for the year, and different celebration dates. In the present booklet we do not encompass the whole variety of these distinct features, and pick up a single aspect, namely, the average length of the year. There are only four acceptably simple and exact options. They are given by the first four columns of the table above. Combinations stemming from the fifth, etc. columns are far too complicated. The possible—and acceptable—options are thus listed in Table 3.

Table 3

Option No.	Alternation of leap years		Average length of a year	Error
	number of leap years	period		
1	1	4	365 d 6 h 00 m 00 s	—11 m 14 s
2	7	29	365 d 5 h 47 m 35 s	+1 m 11 s
3	8	33	365 h 5 h 49 m 05 s	—19 s
4	31	128	365 d 5 h 48 m 45 s	+1 s

The minus sign in the **Error** column indicates that the average duration of the year is greater than the true value.

The first option is the Julian calendar.

The second option is not expedient. It is as complicated as the third option being much less accurate.

The third option (8 leap years in a 33-years cycle) has been proposed by the great Persian and Tajik scholar and poet Omar Khayyam in 1079.

The fourth option is exceptionally accurate. The error of 1 s is of no practical significance. This calendar has therefore been proposed; for example, the Russian astronomer Medler suggested in 1864 to introduce it in Russia from the beginning of the 20th century. It called for *only* one correction to the Gregorian calendar: to jump one leap year every 128 years (i.e. treat this year as an ordinary one). Indeed, the Julian calendar contains 32 leap years per a 128-years cycle.

However, this calendar has been enacted neither in Russia nor anywhere else. The likely reasons are that the 128-years period is not "rounded-off", and that people are strongly accustomed to the existing calendar.

6.2.3. The Secret of Pope Gregory XIII. The preceding Subsection did **not** solve the mystery of Pope Gregory XIII, the Gregorian calendar is not found among the four options in the table. For this reason, having solved the mathematical problem, we shall spend some time with the historical problem. What were the arguments behind the decision of Pope Gregory XIII (or rather, of the commission he had designated)?

Look at a very appealing hypothesis: Pope Gregory XIII leaned to the ratio 31:128 but wanted to replace the period of 128 years with something more convenient, and chose for this 400 years. If 128 years contain 31 leap years, how many are contained in 400 years? The proportion

$$\frac{31}{128} = \frac{x}{400}$$

yields $x = 96.875 \approx 97$. This is precisely the Gregorian calendar: 97 leap years per 400-year cycle.

Quite conclusive, isn't it? Unfortunately, it is wrong.

When arguing out a historical case, including an event in the history of science, we must avoid ascribing our modern way of reasoning to the scolars of yore. Quite the opposite; we ought to try and penetrate their world of ideas and knowledge. Furthermore, speculative arguments of the it-could-quite-likely-have-been-like-this type are not popular with historians. We need to turn up documents and ascertain that "it has been thus and not differently". A very good deal is known about the Gregorian calendar reform, including the scolars who sat on the commission given the job of compiling the project of the reform.

The flaw in our speculative reasoning was this: the duration of the year was not known in the time of Pope Gregory XIII as accurately as we know it nowadays. Gregory XIII's Commission used the astronomical tables compiled by the Academy of Toledo on the order of King Alphonse X (the Wise) of Castile (1221-1284). The length of the year in these tables was

1 year = 365 days 5 hours 49 minutes 16 seconds.

Converted to continued fraction, it gives

$$1 \text{ year} = [365;\ 4,\ 8,\ 7,\ 2,\ 2,\ 17].$$

Its convergents (with the integral part omitted) are

$$\frac{1}{4},\quad \frac{8}{33},\quad \frac{57}{235}.$$

Pope Gregory XIII's Commission thus could not be aware of the ratio $\frac{31}{128}$, whatever method it chose.

It has been already mentioned that the average duration of the year in the Gregorian calendar is 365 days 5 hours 49 minutes 12 seconds, and thus is *27 seconds longer than the true duration.* But this is our attitude, while Pope Gregory XIII was of the opinion that his year was *4 seconds shorter than the true duration.* We thus find that Pope Gregory XIII's Commission could be very much satisfied with the accuracy it has achieved.

We should add that nothing points to the use of continued fractions by the papal commission; continued fractions remained unknown in Europe at the time. Rather, the commission came to its decision by the trial-and-error method. Here is how this could have been done quite easily. According to the Tables of Alphonse X, the Julian year was longer than the true year by 10^m44^s. How many years would it take to accumulate an error of one full day? Divide 24 hours by 10^m44^s:

$$\frac{24^h}{10^m44^s} = \frac{86\,400}{644} \approx 134.$$

It is thus necessary to overlook a leap year once in every 134 years. But this would not be suitable because the coming 134th year may not be a leap year. But $134 \approx \frac{1}{3} \cdot 400$. *Ergo*: overlook three leap years during 400 years. This gives the Gregorian calendar.

Bibliography

We have mentioned in the Preface that this booklet is intended for laymen and presents what can be regarded as the minimum information on continued fractions. The readers who wish to get a broader acquaintance with the subject are recommended the following sources.

1. Moore C. G. *An Introduction to Continued Fractions*; The National Council of Teachers of Mathematics, Washington D. C., 1964. See the bibliography for further reading.

Additional material about the mathematicians mentioned in the present book can be found in:

2. Struik D. J. *A Concise History of Mathematics*, 3rd edition, Dover, New York, 1967.

Printed in the Union of Soviet Socialist Republics