1 The concept of numbers.

In this chapter we will explore the early approaches to counting, arithmetic and the understanding of numbers. This study will lead us from the concrete to the abstract almost from the very beginning. We will also see how simple problems about numbers bring us very rapidly to analyzing really big numbers. In section 7 we will look at a modern application of large numbers to cryptography (public key codes). In this chapter we will only be dealing with whole numbers and fractions. In the next chapter we will study geometry and this will lead us to a search for more general types of numbers.

1.1 Representing numbers and basic arithmetic.

Primitive methods of counting involve using a symbol such as | and counting by hooking together as many copies of the symbol as there are objects to be counted. Thus two objects would correspond to ||, three to |||, four to ||||, etc. In prehistory, this was achieved by scratches on a bone (a wolf bone approximately 30,000 years old with 55 deep scratches was excavated in Czechoslovakia in 1937) or possibly piles of stones. Thus if we wish to record how many dogs we have we would, say, mark a bone with lines, one for each dog. That is 5 dogs would correspond to |||||. Notice, that we are counting by assigning to each dog an abstract symbol for one dog. Obviously, the same method could have been used for cats or cows, etc. Thus the mark | has no unit attached. One can say "|||||| dogs" (dogs being the unit). Notice that you need exactly the same number of symbols as there are objects that you are counting.

Although this system seems very simple, it contains the abstraction of unitless symbols for concrete objects. It uses the basic method of set theory to tell if two sets have the same number of elements. That is, if A and B are *sets* (collections of objects called elements) then we say that they have the same number of elements (or the same *cardinality*) if there is a way of assigning to each element of the set A a unique element of the set B and every element of the set B is covered by this assignment. Primitive counting is done by using sets whose elements are copies of | to be numbers. Although each of the symbols | is indistinguishable from any other they must be considered different. This primitive method of counting and attaching symbols to numbers basically involves identifying sets with the same cardinality with one special set with that cardinality. In modern mathematics, one adds one level of abstraction and says that the set of all sets with the same cardinality constitutes one *cardinal number*. There is no limit to the size of a set in this formalism. We will come back to this point later.

Early methods of representing numbers more concisely than what we have called the primitive system are similar to Roman numerals which are still used today for decorative purposes. In this system, one, two, three are represented by I, II, III. For five there is a new symbol V (no doubt representing one hand) and four is IV (to be considered one before V and probably representing a hand with the thumb covering the palm). Six, seven and eight are given as VI, VII,

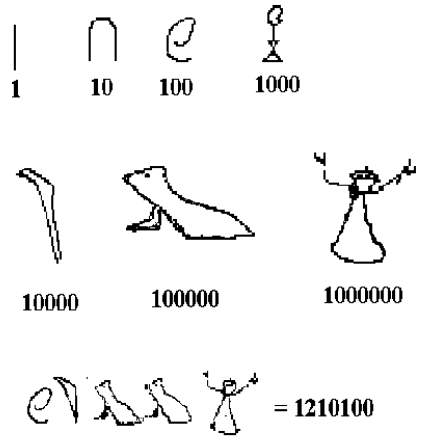


Figure 1:

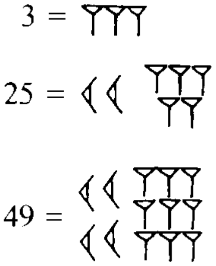
VIII. Then there is a separate symbol for ten, X (two hands) and nine is IX. This pattern continues, so XII is twelve, XV is fifteen, XIV is fourteen, XIX is nineteen. Twenty and thirty are XX, XXX. Fifty is L. Forty is XL. One hundred is C, five hundred is D and a thousand is M. Thus 1998 is MCMXCVIII. This system is adequate for counting (although cumbersome). It is, however, terrible for arithmetic. Here we note that one has a dramatic improvement in the number of symbols necessary to describe the number of elements in a set. Thus one symbol M corresponds to the cardinal with 1000 of the symbols | in it in the most primitive system.

The ancient Egyptians (beginning about 3500 BC) used a similar system except that they had no intermediate symbols for five, fifty or five hundred. But they had symbols for large numbers such as ten thousand, one hundred thousand, one million and ten million. The below is taken from the Rhind Papyrus (about 1600 BC).

Our number system derives from the Arabic positional system which had its precursor in the Babylonian system (beginning about 3000 BC). Before we describe the Babylonian system it is useful to recall our method of writing numbers. We use symbols 1,2,3,4,5,6,7,8,9 for one element, two elements,...,nine elements. we then write 10 for ten elements, 11 for eleven, ..., 19 for nineteen. This means that we count by ones, then by tens, then by hundreds, then by thousands, etc. This way we can write an arbitrarily large number using ten symbols (we also need 0 which will be discussed later). Our system has base ten. That, is we count to nine then the next is ten which is one ten, 10, then we count by ones from 11 to 19 and the next number is two tens, 20. When we get to 9 tens and 9 ones (99) the next number is 10 tens which we write as 100 (hundred). 10 hundreds is then 1000 etc. Thus by hooking together 10 symbols we can describe all numbers.

One could do the same thing using a base of any positive integer. For example, if we worked with base 2 then we would count 1, then 10 for two, then 11 for three (one two and one one), then 100 (2 twos), 101, 110, 111, 1000 (two (two twos)). Thus we would only need 2 symbols in juxtaposition to describe all numbers. For example, 1024 would need 1024 of the units, | ,in the most primitive system, it is 4 symbols long in ours, and base 2 it is 10000000000. Still a savings of 1013 symbols. The Roman method would be MXXIV so in this case slightly worse than ours. However, if we try 3333 in Roman notation we have MMMCCCXXXIII. How long is the expression for 3333 in base 2?

The Babylonians used base 60 which is called *sexagesimal*. We should note that for some measurements we still use this system: 60 seconds is a minute, 60 minutes is an hour. Their system is preserved in clay tablets in various excavations. Their method of writing (cuneiform) involved making indentations in soft clay tablets by a wedge shaped stylus.



They used two basic symbols, one equivalent with | for one. and one for 10 which we will represent as ◁. Thus six is ||||||. Normally written in the form:

$$\begin{matrix} ||| \\ ||| \end{matrix}$$

and thirty seven is

$$\begin{array}{ll} \triangleleft & ||| \\ \triangleleft & |||\ . \\ \triangleleft & \ | \end{array}$$

But 61 is | |. 3661 is | | |. Thus, except that they used only symbols for 1 and 10 and had to juxtapose them to get to 59, they used a system very similar to ours. They did not have a symbol for 0. We will see that this is a concept that would have to wait more than 3000 years. So when they saw |, they would have to deduce from the context whether it represented 1, 60, 3600, etc. For example if I said that a car cost ||| then you would be pretty sure (in 2003) that I meant 10,800, not 180 or 3. They later (200 BC) had a symbol that they could use for a place marker in all but the last digit (but still no 0). // Thus they could write |//| and mean 3601. There is still an ambiguity in the symbol | which can still mean 1, 61, 3601, etc.

Exercises. 1. Write out the number 1335 in Egyptian notation, binary, sexagesimal and in Roman numerals.

2. For computers one kilobit (1K) is actually 1024. Why is that?

3. The early computer programmers used base 16 they therefore needed 16 symbols which they wrote as 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F. For example,$AF = 10 \times 16 + 15 = 175$. What number is $FFFF$? Write it in binary. Why was it important to 16 bit computers? $FFFFF + 1$ is called a megabit. Why is that?

4. In writing numbers in the Egyptian system what is the maximum repetition necessary for each symbol?

1.2 Arithmetic.

1.2.1 Addition.

We return to the most primitive method of counting. If you have ||| sheep and you have purchased |||| sheep, then you have ||||||| sheep. That is, to add ||| and |||| we need only "hook" |||| onto |||. For cardinal numbers we have thus described a method of addition: If A corresponds (i.e. is an element of) to the cardinal a and if B corresponds to the cardinal b, and if no element of A is an element of B then $a + b$ is the cardinal number that contains $A \cup B$ (with $A \cup B$ the set that consists of the elements of A combined with those of B).This can be made rigorous (independent of the choice of A and B) we will look into this point later in the book. Thus the abstraction of primitive addition is set theoretic *union* of *disjoint* (no element in common) sets.

In the Roman system there is one more degree of abstraction since for example |||| is represented as IV and ||||| is represented as V so IV + V = ||||||||| = IX. Obviously, one must remember much more if one uses the more abstract method of the Romans than the direct "primitive" method.

In our system for the same addition we are looking at $4 + 5$ and we must remember that this is 9. Thus the situation is analogous to that of the Romans. However, if we wish to add XXXV to XVI, then in Roman numerals we have LI. In our system we have $35 + 16$. We add $5 + 6$ and get 11 (memorization). We now know that the number has a 1 in the "ones position" we carry the other 1 to see that for the "tens position". We have $1 + 3 + 1 = 5$. The sum is therefore 51. Thus we need only remember how to add pairs of numbers up to 9 in our system and all other additions are done following a prescribed method. The Roman system clearly involves much more memorization.

We next look at the Babylonian system. For this we will use a method of expressing numbers to base 60 that is due to O. Neugebauer(a leader in the history of mathematics). We write 23,14,20 for 20 plus 14 sixties plus 23 $\times 3600$. Thus in the Babylonian base 60 system we must memorize all additions of numbers up to 59. If we wish to add 21,13 and 39,48 then we add $48 + 13$ and get 1,1 (this is memorized or in an addition table) 21+39 and get 1 (remembering the context). Thus the full sum is 1,1,1. Here we must remember a very large addition table. However, we have grown up thinking in terms of base 10 and we do the additions of pairs of numbers below 59 in our method and then transcribe them to our version of the Babylonian notation.

Exercises.

1. Do the addition $1,2 + 32,21,3 + 43,38,1$ in Neugebauer's notation.
2. How do you think that an Egyptian would add together 3076 and 9854?

1.2.2 Multiplication.

Multiplication is a more sophisticated operation than addition. There isn't any way to know when and how the notion arose. However, the Egyptians and the Babylonians knew how to multiply (however as we shall see the Egyptian method is not exactly what one would guess). We understand multiplication as repeated addition. That is, if we wish to multiply a times b, $a \cdot b$, then we add b to itself a times. That is $3 \cdot 5$ is $5 + 5 + 5 = 15$.

If we attempt to multiply a times b in the primitive system we must actually go through the full juxtaposition of b with itself a times (or vice-versa). In a system such as the Roman system we must memorize a great deal. For example XV·LI = DCCLXV. For us the multiplication is done using a system:

$$\begin{array}{r} 51 \\ \underline{15} \\ 255 \\ \underline{510} \\ 765 \end{array} .$$

We usually leave out the 0 in the 510 and just shift 51 into the position it would have if there were a 0. We see that we must memorize multiplication of pairs of numbers up to 9.

The Babylonian system is essentially the same. However, one must memorize multiplication of pairs of numbers up to 59. This is clearly a great deal to remember and there are tablets that have been excavated giving this multiplication table.

The Egyptian system is different. They used the method of *duplication.* For example if we wish to multiply 51 by 15 then one would proceed as follows:

$$\begin{array}{cc} 51 & 1 \\ 51 + 51 = 102 & 2 \\ 102 + 102 = 204 & 4 \\ 204 + 204 = 408 & 8 \end{array}$$

Now $1 + 2 + 4 + 8 = 15$ so the product is $51 + 102 + 204 + 408 = 765$. Notice that they are actually expanding 15 in base 2 as 1111. If the problem had been multiply 51 by 11 then the answer would be $51 + 102 + 408 = 561$ (in base 2, 11 is 1011). So their multiplication system is a combination of doubling and addition.

We note that this method is used in most computers. Since, in base 2, multiplication by 2 is just putting a 0 at the end of the number. In base 2,

$51 = 110011$. Thus the same operations are

110011	1
1100110	10
11001100	100
110011000	1000

The basic difference is that we must remember many carries in binary. Thus it is better to proceed as follows.

$$\begin{array}{r} 110011 \\ \underline{1100110} \\ 10011001 \end{array}$$

$$\begin{array}{r} 11001100 \\ \underline{110011000} \\ 1001100100 \end{array}$$

$$\begin{array}{r} 10011001 \\ \underline{1001100100} \\ 1011111101 \end{array}$$

Addition is actually an operation that involves adding *pairs* of numbers. In our system we rarely have to carry numbers to more than one column to the left (that is when a column adds up to more than 99). In binary it is easy if we add 4 numbers with 1 in the same digit we will have a double carry.

Exercise.

1. Multiply 235 by 45 using the Egyptian method. Also do it in binary.

1.2.3 Subtraction.

If we wish to subtract $|||$ from $|||||$ then the obvious thing to do is to remove the lines one by one from each of these primitive numbers. $|||\ ||||| \to ||\ |||| \to |\ ||| \to ||$. With this notion we have $||||| - ||| = ||$. If we do this procedure to subtract a from b and run out of tokens in b then we will say that the subtraction is not possible. This is because we have no notion of negative numbers. We will see that the concept of 0 and negative numbers came relatively late in the history of mathematics. In any event, we will write $a < b$ if subtraction of a from b is possible. If $a < b$ then we say that a is strictly less than b.

In our notation subtraction is an inverse process to addition. This is because our number notation has a higher degree of abstraction than the primitive one. Thus we memorize such subtractions as $3 - 2 = 1$. If we are calculating $23 - 12$ then we subtract 2 from 3 and 1 from 2 to get 11. For $23 - 15$ we do in initial borrow and calculate $13 - 5$ and $1 - 1$. So the answer is 8. Obviously, we can only subtract a smaller number from a larger one if we expect to get a number in the sense we have been studying. Both the Egyptians and the Babylonians used a similar system. For the Egyptians it would be somewhat more complicated, since every new power of 10 entailed a new symbol.

1.2.4 Division and fractions.

For us division is the inverse operation to multiplication in much the same way as subtraction is the inverse operation to addition. Thus $\frac{a}{b}$ is the number such that if we multiply it by b we have a. Notice that if b is 0 this is meaningless and that even if a and b are positive integers then $\frac{a}{b}$ is not always an integer. Integer division can be implemented as repeated subtraction thus in the primitive notation |||||||| - ||| = |||||, ||||| - ||| = ||| thus ||||||||/||| = |||. However, the Egyptians and Babylonians understood how to handle divisions that do not yield integers.

The ancient Egyptians created symbols for the fractions $\frac{1}{n}$ (i.e. reciprocals). They also had a symbol for $\frac{2}{3}$. However, if they wished to express, say, $\frac{7}{5}$ then they would write it as a sum of reciprocals say $1+\frac{1}{3}+\frac{1}{15}$. Also they limited their expressions to *distinct* reciprocals (or $\frac{2}{3}$). Thus $1+\frac{1}{5}+\frac{1}{5}$ was not a valid expression. Note that such an expression is not unique. For example, $\frac{7}{5}=1+\frac{1}{4}+\frac{1}{10}+\frac{1}{20}$. Notice that one allows any number of reciprocals in the expression. With a method such as this for handling fractions, there was a necessity for tables of fractions. One also had to be quite ingenious to handle fractions. An ancient Egyptian problem asks:

If we have seven loaves of bread to distribute among 10 *soldiers, how would we do it?*

We would instantly say that each soldier should get $\frac{7}{10}$ of a loaf. However, this makes no sense to the ancient Egyptians. Their answer was (answers were supplied with the problems) $\frac{2}{3}+\frac{1}{30}$.

The mathematician Leonardo of Pisa (Fibonacci 1175-1250 A.D.) devised an ingenious method of expressing fractions in the Egyptian form. In order to see that the method works in general several basis properties of numbers will be used here. They will be considered in more detail later He starts by observing that we need only consider $\frac{a}{b}$ with $1 < a < b$. We first observe that $\frac{b}{a} > 1$ so there exists a positive whole number n such that

$$n-1 < \frac{b}{a} < n.$$

So

$$\frac{1}{n} < \frac{a}{b} < \frac{1}{n-1}$$

Thus $\frac{a}{b}=\frac{1}{n}+\frac{an-b}{nb}$. We observe that $an-b>0$ and $a-(an-b)=b-a(n-1)=a(\frac{b}{a}-n-1)>0$. Thus $a>(an-b)>0$. Set $a_1=an-b$, $b_1=bn$. Then $0<a_1<a$ and $1\le a_1<b_1$. If $a_1=1$ then we are done. Otherwise, we repeat the process with $\frac{a_1}{b_1}$. Call n, n_1. If $a_1=1$ then we see that $b_1=n_1b>n_1$ so $\frac{a}{b}=\frac{1}{n_1}+\frac{1}{n_1b}$ is a desired expression. Assume that $a_1>1$. Do the same for $\frac{a_1}{b_1}$. This is Fibonacci's method. A full proof that this always works didn't get published until the nineteenth century and is attributed to J.J.Sylvester. What has not been shown is that if

$$\frac{1}{n} < \frac{a_1}{b_1} < \frac{1}{n-1}$$

then $n > n_1$. We do this by showing that $\frac{a_1}{b_1} < \frac{1}{n_1}$. To see this we observe that $an_1 - b < a < b$. Thus

$$\frac{a_1}{b_1} = \frac{an_1 - b}{n_1 b} < \frac{a}{n_1 b} < \frac{b}{n_1 b} < \frac{1}{n_1}.$$

In this argument we used an assertion about not necessarily whole numbers that says that if we have a number then it lies between two consecutive integers.

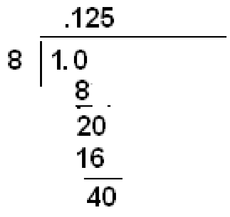
Consider, for example, $\frac{7}{10}$ then $n_1 = 2$ and $\frac{7}{10} - \frac{1}{2} = \frac{1}{5}$. Thus we get as an answer to the Egyptian problem $\frac{1}{2} + \frac{1}{5}$ which seems preferable to the answer given in the original Papyrus.

Fibonacci was one of the leading European mathematicians of the Middle Ages. He was instrumental in introducing the Arabic number system (the one we use) to the West. However, he preferred the Egyptian method of fractions to our decimal notation (below). Clearly one must be much cleverer to deal with Egyptian fractions than with decimals. Also, as we will see, strange and impractical problems have propelled mathematics to major new theories (some of which are even practical).

The ancient Babylonians used a method that was analogous to our decimal notation. In our decimal method we would express a fraction such as $\frac{1}{8}$ as follows: We first try to divide 8 into 1 this fails so we multiply by 10. We can divide 8 into 10 once with remainder 2. We must multiply by 10 and divide 8 ani 20 getting 2 with remainder 4. We now multiply 4 by 10 and get 40. Divide by 8 and get 5. The numbers for the three divisions are $1, 2, 5$. We write

$$\frac{1}{8} = .125.$$

We can express this as follows:



We will use Neugebauer notation in our description of their method. The fraction $\frac{7}{5} = 1 + \frac{24}{60}$. Our version of their notation would be $1;24$. In our decimal notation this is 1.4. We could use exactly the same process (though it is harder for us to do the intermediate steps in our heads). We must divide 5 into 2. So we multiply by 60 and do the division. That is divide 120 by 5. We get 24 and no remainder. If we were to write $\frac{1}{8}$ we could work as follows:1×60 divided by 8 is 7 with remainder 4. $4 \times 60 = 240$ which divided by 8 is 30 with no remainder. Thus we have $;7,30$.

There were also bad fractions. In our decimal notation

$$\frac{1}{3} = 0.33333...$$

That is we must write the symbol 3 forever. In the Babylonian form $\frac{1}{7}$ is the first bad fraction and it is given by

$$;8,34,17,8,34,17,...$$

repeating 8,34,17 forever. Suddenly the Egyptian way doesn't seem to be so silly!

We also think of fractions as expressions $\frac{p}{q}$ with p and q positive integers. $\frac{p}{q} = \frac{r}{s}$ means $ps = rq$. Addition is given by $\frac{p}{q} + \frac{r}{s} = \frac{ps+qr}{qs}$. Multiplication is given by $\frac{p}{q} \cdot \frac{r}{s} = \frac{pr}{qs}$. We note that $\frac{1}{2} = \frac{2}{4} = \frac{3}{6} = ...$ That is we identify all of the symbols $\frac{n}{2n}$ with $\frac{1}{2}$. This is similar to our definition of cardinal number. Usually, to rid ourselves of this ambiguity, we insist that p and q are in lowest terms. That is they have no common factor other than 1

1.2.5 Exercises.

1. Why do you think that the Egyptians preferred $\frac{2}{3} + \frac{1}{30}$ to $\frac{1}{2} + \frac{1}{5}$ for $\frac{7}{10}$?

2. Use the Fibonacci method to write $\frac{4}{17}$ as an Egyptian fraction.
3.Make a table in Egyptian fractions of $\frac{n}{10}$ for n equal to $1,2,3,4,5,6,7,8,9$

4. Among $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, ..., \frac{1}{19}\}$ which are the bad fractions in base 10? What do they have in common. Can you guess a property of n that guarantees that $\frac{1}{n}$ is a good fraction to base 10? How about base 60?

5. The modern fame of Fibonacci is the outgrowth of a problem that he proposed: Suppose that it takes a rabbit exactly one month from birth before it is sexually mature and that a sexually mature pair (male and female) of rabbits will give birth to two rabbits (male and a female) every month. If we start with newly born male and female rabbits how many rabbits will there be at the end of one year? What is the answer to his question?

6. If b is a positive integer then we can represent any integer to base b in the form $a_0 + a_1 b + a_2 b^2 + ... + a_k b^k$ with $0 \leq a_i < b$.. This if $b = 10$ then 231 means $1 + 3 \times 10 + 2 \times 10^2$. If $b = 60$ then 231 means $1 + 3 \times 60 + 2 \times 60^2$. Show that if $n < b$ then the square of 111...1 (n ones) is 123...n...321 that is the digits increase to n then decrease to 1. What happens if $n > b$?

Three examples of early Algebra. At this point we have looked at counting methods and developed the basic operations of arithmetic. We have studied one simple Egyptian exercise in arithmetic and given a method of Fibonacci to express a fraction as an Egyptian fraction. The start with a "practical problem" or applied mathematics. Whereas, by the time Fibonacci devised his method, there was no reason to use Egyptian fractions. It is what we now call pure mathematics. The method is clever and has an underlying simplicity that is much more pleasant than using trial and error. Obviously, ancient peoples had many uses for their arithmetic involving counting, commerce, taxation, measurement, construction, etc. But even in the early cultures there were mathematical puzzles and techniques developed that seem to have no practical use.

An Egyptian style problem:

A quantity added to two thirds of it is 10. *What is the quantity*?

We would say set the quantity equal to x (we will see that this small but critical step would not be discovered for thousands of years). Then we have

$$x + \frac{2}{3}x = 10.$$

Hence

$$\frac{5}{3}x = 10.$$

So $x = 6$. Since the Egyptians had no notion of how to deal with unknown quantities, they would do something like. If the quantity were 3 then the sum of the quantity plus two thirds of the quantity is 5. Since the sum we desire is 10, the answer is 2 times 3 or 6. In other words, they would use a convenient value for the quantity and see what the rule gave for that value. Then re-scale to get the answer.

We will now discuss a Babylonian style problem (this involves basic geometry which we will assume now and discuss in context later). Before we write it out we should point out that multiplication as repeated addition was probably not an important motivation for doing multiplication. More likely they multiplied two numbers because the outcome is the area of the rectangle whose sides were the indicated number of whatever units they were using.

I add the area of a square to two thirds of its side and I have ;35. *What is the side of the square?*

Solution:

Take 1 *multiply by* $\frac{2}{3}$ *take half of this and we have* $;20$. *You multiply this by* $;20$ *and the result is* $;6,40$. *You add to this ;35 to have* $;41,40$. *This is the area of the square of side* $;50$. *You subtract* $;20$ *from* $;50$ *and you have* $;30$ *the side of the square.*

In our notation what we have done is taken $\frac{2}{3}$. Next divided by 2 to get $\frac{1}{3}$. The square of $\frac{1}{3}$ is $\frac{1}{9}$ now add $\frac{35}{60} = \frac{7}{12}$ to get $\frac{25}{36}$. The square root of this is $\frac{5}{6}$. Subtract $\frac{7}{12}$ and we have $\frac{1}{2}$. Thus if $a = \frac{2}{3}$, $b = \frac{7}{12}$ then the answer is

$$\sqrt{\left(\frac{a}{2}\right)^2 + b} - \frac{a}{2}.$$

In modern notation if we set the side equal to x then we are solving

$$x^2 + \frac{2}{3}x - \frac{7}{12} = 0.$$

The quadratic formula tells us that if we are solving

$$x^2 + ax - b = 0$$

then

$$x = \frac{-a \pm \sqrt{a^2 + 4b}}{2}.$$

If $a > 0$,$b > 0$ then the positive solution is exactly the Babylonian answer. Their method of solving such problems put a premium on the ability to calculate expressions of the form $\sqrt{a^2 + b}$. They had an approximate method of doing such calculations which corresponds to what we will see is the second iteration of a method of Newton method applied to this simple case. They use the approximation $a + \frac{b}{2a}$. Notice that if $\frac{b}{2a}$ is small then this is a good approximation.

Thus the Babylonians were aware of general methods to solve quadratic equations. They, however, could only express their method in words. What they wrote out is except for the order (and the absorption of the $\frac{1}{2}$) exactly what we would write.

It is hard to imagine how either of these methods could be used in practical applications. However, one of the most interesting exercises in pure mathematics can be found in a tablet in the Yale collection (Plimpton 322). This tablet is a tabulation of 15 triples of numbers a, b, c with the property that

$$a^2 + b^2 = c^2.$$

The simplest example that we know of this is

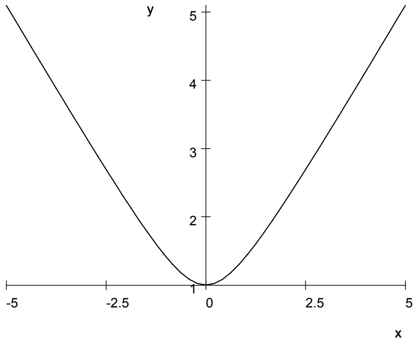
$$3^2 + 4^2 = 5^2.$$

This triple appears on the tablet as number 11 and in the form

$$60^2 + 45^2 = 75^2.$$

The tablet is thus using some strange rule for generating these numbers (usually called *Pythagorean triples*). We will discuss the Pythagorean theorem later. Here we will study the tablet as a collection of relationships between numbers. The table is arranged as follows: there are 15 existent rows and 4 readable columns. The first column contains a fraction and the fractions are decreasing.

The second and third contain integers and the last is just the numbers 1,...,15 in order. If we label an element of the second column a and the element of the third column in the same row c then $c^2 - a^2 = b^2$ with b a positive integer and the element of the first column in the same row is $\frac{c^2}{b^2}$. Also the first column contains only regular sexagesimal rational numbers. It seems clear that the Babylonians were aware of a method of generating Pythagorean triples.

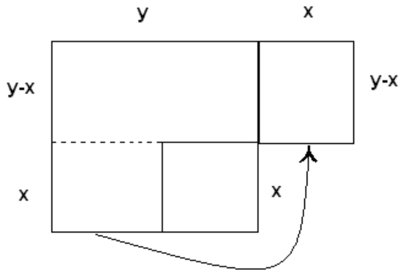


In our modern notation we know how to generate all Pythagorean triples a, b, c $(a^2 + b^2 = c^2)$ with a,b,c having no common factor. Indeed, consider $y = \frac{c}{b}$, $x = \frac{a}{b}$ then $y^2 - x^2 = 1$. We are thus looking for rational points on a hyperbola (see the figure above). Notice that $\frac{1}{x^2}$ gives an element of the first column of the table. Thus they seem to have picked points rational points on the hyperbola in increasing order. How do you locate such a point?

We note that $y^2 - x^2 = (y - x)(y + x)$ (we will discuss what this might have meant to the Babylonians soon). We write $y + x = \frac{m}{n}$, $y - x = \frac{n}{m}$ then $y = \frac{1}{2}\left(\frac{m}{n} + \frac{n}{m}\right)$ and $x = \frac{1}{2}\left(\frac{m}{n} - \frac{n}{m}\right)$. Thus $y = \frac{m^2+n^2}{2mn}$ and $x = \frac{m^2-n^2}{2mn}$. This suggests that we take $a = m^2 - n^2$, $b = 2mn$ and $c = m^2 + n^2$. If m and n are positive integers then you can check easily that this assignment generates a Pythagorean triple. There is an algebraic proof of Fibonacci that this method generates all Pythagorean triples that have no common factor. André Weil (1906-1999) has pointed out that there is a geometric argument in Euclid Book X, Lemma 1,2 in preparation for Proposition 29 that proves that this method gives all such triples that are relatively prime (in fact a bit more than this). We will come back to this later.

Consider the Pythagorian triple $3, 4, 5$. We will find numbers m, n as above. The method above says take $y = \frac{5}{4}$ and $x = \frac{3}{4}$. Then $y + x = 2$ and $y - x = \frac{1}{2}$. This suggests take $m = 2$ and $n = 1$. We can check that this works $m^2 - 1 = 3$, $2mn = 4$ and $m^2 + n^2 = 5$.

We will now discuss a probable meaning for the formula $y^2 - x^2 = (y - x)(y + x)$. The formula $y^2 - x^2$ is geometrically the area of the figure gotten by removing a square of side x from one of side y. If you take the smaller square out of the lower right corner then in the lower left corner one has a rectangle of side x and base $y-x$. If we cut this rectangle off and rotate it 90^o then we can attach it to what is left of the big square and get a rectangle of side $y-x$ and base $y+x$.



The two problems above are similar to the "word problems" of high school algebra and were probably used in the same way as we use them now. That is, to hone the skills of a student learning basic algebra. Plimpton 322 is another matter. It contains number theoretic relationships at a sophisticated level. Imagine a line of reasoning similar to the one in the previous paragraph without any algebraic notation and without even the notion of a fraction.

1.2.6 Exercises.

1. Problem 26 on the Rhind papyrus is:
A quantity whose fourth part is added to it becomes 15.
Use the Egyptian method to solve the problem.

2. Use the Babylonian approximation to calculate $\sqrt{2}$. (Suggestion: Start with $a=\frac{4}{3}$ so that $b=\frac{2}{9}$. Can you improve on this?)

3. A problem on a Babylonian tablet says:
I have added 7 times the side of my square to 11 times the area and have $6;15$*. Find the side.*
Use the Babylonian method to solve this problem.

4. Find m,n so that $a=m^2-n^2$, $b=2mn$ and $c=m^2+n^2$ for the Pythagorean triples $119,120,169$ and $5,12,13$.

1.3 Some number theory taken from Euclid.

We now jump about 1500 years to about 300BC and the time of the school of Euclid in Alexandria. We will examine parts of Books VII,VIII,IX of his *Elements* that deal with numbers. We will have more to say about the other books at appropriate places in this work. We will use the translation of Sir Thomas Heath for our discussion.

1.3.1 Definitions

Euclid begins Book VII with 22 definitions that set up basic rules for what we have been calling the primitive number system. We will see in the next chapter that Euclid did not think of numbers in this sense. He rather thought of numbers as intervals. If we have two intervals AB and CD and if we lay out AB a certain number of times an this covers CD exactly then AB is said to *measure* CD.

1. An *unit* is that by virtue of which each of the things that exist is called one.

This doesn't make too much sense but it is basically establishing that there is a unit for measurement.. We have been denoting this by |.

2. A *number* is a multitude composed of units.

Thus ||| is a number as before. However, Euclid thinks of it as an interval that is exactly covered by three unit intervals.Be warned that the unit is not considered to be a number.

3. A number is *a part* of a number, the less of the greater, when it measures the greater;

Thus the greater, ||||||, is measured by the less |||.

4. but parts when it does not measure it.

||||| is not measured by |||.

5. The greater number is a *multiple* of the less when it is measured by the less.

Notice that the definitions are beginning to be more accessible. Here we measure |||||| by two of the |||. This thus |||||| is ||| multiplied by ||.

6. An *even number* is that which is divisible into two equal parts.

7. An *odd number* is that which is not divisible into two equal parts, or that which differs by a unit from an even number.

8.,9.,10. talk about multiplication of odd and even numbers. (e.g. an odd by an even is an even).

11. A *prime number* is that which is measured by a unit alone.

Thus |||||| is measured by |, ||, ||| so is not prime. ||||| is only measured by |.

12. Numbers *prime to one another* are those which are measured by an unit as a common measure.

|||| is measured by |, || ||||||||| is measured by |, ||| thus the only common measure is |. Thus |||| and ||||||||| are prime to one another.

13., 14. are about numbers that are not prime (to each other). A number that is not prime is *composite*.

In 15. he describes multiplication as we did (repeated addition).

16. And when two numbers having multiplied one another make some number, the number so produced is called *plane*, and its *sides* are the numbers which have been multiplied.

Here Euclid seems to want to think of the operation of multiplication in geometric terms: an area.

In 17 the product of three numbers is looked upon as a *solid*.

18,.19. define a *square* and a *cube* as we do. We will study these concepts in the next chapter.

20. Numbers are *proportional* when the first is the same multiple, or the same part, or the same parts, of the second that the third is to the fourth.

||| |||||| and |||| |||||||| are proportional. This is a relationship between two pairs of numbers. It is essentially our way of looking at rational numbers.

In 21. there is a discussion of similar plane and solid numbers.

22. A *perfect number* is one which is equal to its parts.

The parts of |||||| are |, ||, ||| and $| + || + ||| =$ ||||||. So it is perfect. |||| is not.

To us this is not a very basic concept. Perfect numbers are intriguing (28 is one,what is the next one?) but it is hard to see any practical reason for their study. We shall see that Euclid gave a method for generating perfect numbers.

1.3.2 Some Propositions

Having disposed of the definitions, Books VII,VIII,IX consist of a series of Propositions. Number one is:

*B*_____________*F*__*A*

*D*________*G*__*C*

__*E*

Two unequal numbers being set out, and the less being continually subtracted in turn from the greater, if the number which is left never measures the one before it until an unit is left, the original numbers will be prime to one another.

Let us try this out. Take 27 for the larger and 8 for the smaller. Subtract 8 from 27 and get 19, subtract 8 and get 11, subtract 8 and get 3, subtract 3 from 8 and get 5 subtract 3 from 5 and get 2 subtract 2 from 3 and get 1. Thus the numbers are relatively prime.

We will now describe the Euclidian proof. The numbers are denoted AB and CD and Euclid draws them as vertical intervals. He assume on the contrary that AB and CD are not prime to each other. Then there would be a number E that measures both of them. We now come to the crux of the matter: "Let

CD measuring BF leaving FA less than itself." (Here it is understood that $BF + FA = BA$ and that BF is evenly divisible by CD)

This assertion is now called the *Euclidean algorithm*. It says that if m, n are whole numbers with $m < n$ then we can write $n = dm$ or $n = dm + q$ with q a whole number strictly less than m. For some reason he feels that this assertion needs no proof. To Euclid this is evident. If n is measured by m it is obvious. If it is not then subtract m from n and get q_1 if $q_1 < m$ we are done. q_1 cannot be measured by m hence $q_1 \neq m$ and so $q_1 > m$. We now subtract m from q_1 and get q_2. If $q_2 < m$ then we are done otherwise as before $q_2 > m$. Subtract m again. This process must eventually lead to a subtractend less than m since if not then after n steps we would be able to subtract nm from n so $mn < n$. But this is impossible since $m > 1$ so $mn = n + n + ... + n$ (m times). Hence we are asserting $n > mn \geq n + n$. Since it is obvious that $n + n > n$ we see that the process must give the desired conclusion after less than n steps.

We now continue the proof. Let AF measuring DG leaving GC less than itself. E measures CD hence BF and E measures AB so E measures FA. Similarly, E measures GC. Since the procedure described in the proposition now applies to AF and GC, we eventually see that E will eventually measure a unit. Since E has been assumed to be a number (that is made up of more that one unit) we see that this is impossible. In Euclid this unbounded procedure (finite for each example) is only done three times. Throughout his arguments he does the case of three steps to represent the outcome of many steps.

The second proposition is an algorithm for calculating the greatest common divisor (greatest common measure in to Euclid).

Given two numbers not prime to one another, to find the greatest common measure.

Given AB and CD not prime to one another then and CD the smaller then if CD measures AB then it is clear that CD is the greatest common measure. If not consider $AB - CD$, CD. There are now two possibilities. The first is that $AB - CD$ is smaller than CD. In this case if $AB - CD$ measures CD then it must measure AB and so is the greatest common measure. In the second case CD is the smaller and if CD measures $AB - CD$ then it must measure AB and so it is the greatest common measure. If not the previous proposition implies that if we continually subtract the smaller from the larger then we will eventually come to the situation when the smaller measures the larger. We thus have the following procedure: we continually subtract the smaller from the larger stopping when the smaller measures the larger. Proposition 1 implies that the procedure has the desired end.

Here is an example of proposition 2. Consider 51 and 21. Then $51 - 21 = 30$ (30,21) , $30 - 21 = 9$(21,9), $21 - 9 = 12$ (12,9), $12 - 9 = 3$ (9,3) so the greatest common divisor is 3.

Why is it so important to understand the greatest common divisor? One important reason is that it is the basis of understanding fractions or rational numbers. Suppose that we are looking at the fraction $\frac{21}{51}$. Then we have seen that the greatest common divisor of 21 and 51 is 3. Dividing both 21 and 51 by 3 we see that the fraction is the same as $\frac{7}{17}$. This expression is in *lowest terms* and is unique. $\frac{7}{17} = \frac{21}{51} = \frac{42}{102} = ...$

We will emphasize his discussion of divisibility and skip to Proposition 31.

Any composite number is divisible by some prime number.

We will directly quote Euclid.

Let A be a composite number; I say that A is measured by some prime number. For since A is composite, some number will measure it. Let a number measure it, and let it be B. Now, if B is prime then we are done. If it is composite then some number will measure it. Let a number measure it and call it C. Since C measures B and B measures A, C measures A. If C is prime then we are done. But if it is composite then some number will measure it. Thus, if the investigation is continued in this way, some prime number will be found which will measure the number before it, which will also measure A. For if it were not found an infinite series of numbers will measure the number, A, which is impossible in numbers.

Notice that numbers are treated more abstractly as single symbols A, B, C and not as intervals. (Although they are still pictured as intervals.) More important is the "infinite series" of divisors of A. No real indication is given about why this is impossible for numbers. However, we can understand that Euclid considered this point obvious. If D is a divisor of A and not equal to A then D is less than A. There are only a finite number of numbers less than a given number $n, 1, 2, ..., n-1$. This argument uses a version of what is now called mathematical induction which we will call the *method of descent.* Suppose we have statements P_n labeled by $1, 2, 3, ...$ If whenever P_n is assumed false we can show that there is an m with $1 < m < n$ with P_m false then P_n is always true. The proof that this method works is that if the assertion for some n were false then there would be $1 < m_1 < n$ for which P_{m_1} is false. But then there would be $1 < m_2 < m_1$ for which P_{m_2} is false and this procedure would go on forever. Getting numbers $m_1 > m_2 > ... > m_n > ...$ with all the numbers bigger than 1.

Let us try in out. The assertion P_n is that if n is not a unit then n is divisible by some prime. If P_n is false that n is not a prime and not a unit. Hence it is composite so it is a product of two numbers a, b neither of which is a unit and both less than n. If P_a were true then a would be divisible by some prime. But that would imply that n is divisible by some prime. This is contrary to our assumption. Thus if P_n is false then P_a is false with $1 < a < n$. The method of descent now implies that P_n is true for all n.

We will now jump to Book IX and Proposition 20.

Prime numbers are more than any assigned multitude of prime numbers.

Here we will paraphrase the argument. Start with distinct primes A,B,C. Let D be the least common multiple of A, B, C (this has been discussed in Propositions 18 and 19 of Book IX). In modern language we would multiply them together. Now consider $D + 1$. If $D + 1$ were composite then there would be a prime E dividing it. If E were one of A,B,C then E would divide 1.

Notice that we are back with three taking the place of arbitrarily large. The modern interpretation of this Proposition is that there are an infinite number of primes. What is really meant is that if the only primes are p_1,...,p_k then we have a contradiction since $p_1 \cdots p_k + 1$ is not divisible by any of the primes and this contradicts the previous proposition.

After Book IX, Proposition 20 there are Propositions 21-27 that deal with combining even and odd numbers and seem to be preparatory to Euclid's method of generating perfect numbers. For example, Proposition 27 (in modern language) says that if you subtract an even number from an odd number then the result is an odd number. Here one must be careful and also prove that if you subtract an odd number from an even number you get an odd number (Proposition 25). We would say that the two statements are essentially the same since one follows from the other by multiplication by -1. However, since negative numbers were not in use in the time of Euclid Proposition 25 and 27 are independent.

We now record one implication of Proposition 31 (and Proposition 30 which is discussed below) that is not explicit in Euclid (we will see why in the course of our argument). This Theorem is usually called the *fundamental theorem of arithmetic.*

If A is a number (hence is not a unit) then A can be written uniquely (up to order) in the form $p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$ with $p_1, ..., p_r$ distinct primes and $e_1, ..., e_r$ numbers (here B^m is B multiplied by itself m times).

We first show that any number is a product of primes using a technique analogous to the method of Euclid in his proof of Proposition 31. If A is prime then we are done. Otherwise A is composite hence by Proposition 31 $A = q_1 A_1$ with A_1 not a unit and q_1 a prime. If A_1 is a prime then we are done. Otherwise, $A_1 = q_2 A_2$ with q_2 a prime and A_2 not a unit. If A_2 is prime we are done since then $A = q_1 q_2 A_2$. Otherwise we continue this procedure and either we are done in an a finite number of steps or we have $A_1 > A_2 > ... > A_n > ...$ an infinite sequence of positive numbers. This is impossible for numbers according to Euclid. We have mentioned in our discussion of Proposition

Let us show how the principle of descent can be used to prove the assertion that every number is a product of primes. Let P_n be the assertion that if n is not the unit then is a product of primes. If P_n is false then n is not a unit and not prime so n is composite. Hence $n = ab$ with neither a nor b a unit. If both P_a and P_b were true then a is a product of primes and b is a product of primes so ab is a product of primes. Thus one of P_a or P_b would be false. If

P_a is false set $m = a$ otherwise P_b is false and set $m = b$. Then $1 < m < n$ and P_m is false. The principle implies that P_n is true for all n.

This principle can be made into a direct statement which we call the *principle of mathematical induction.*. The idea is as follows if $S_1, S_2, ...$ are assertions and if S_1 is true and if the truth of S_m for all $1 < m < n$ implies that S_n is true then S_n is true for all n. This is intuitively clear since starting with S_1 which we have shown is true we have. S_1 is true so S_1 and S_2 are true so S_3 is true, etc. For example suppose that the statement S_n is the assertion

$$1 + 2 + ... + n = \frac{n(n+1)}{2}.$$

Then S_1 says that $1 = 1$ which is true. We now assume that S_m is true for all $1 < m < n$. Then

$$1 + 2 + ... + (n-1) + n = (1 + 2 + ... + (n-1)) + n = \frac{n(n-1)}{2} + n =$$

$$\frac{n(n-1)}{2} + \frac{2n}{2} = \frac{n(n+1)}{2}$$

which is the assertion S_n.

Let us see how the method of descent implies the principle of mathematical induction. Suppose we have a statement S_n for $n = 1, 2, ...$ and suppose that we know that S_1 is true and whenever we assume S_m is true for $1 < m < n$ the S_n is true. Assume that S_n is false. Then n cannot be the unit. If S_m were true for all $1 < m < n$ then we would know that S_n were true. Since we are assuming the contrary we must have S_m is false for some m with $1 < m < n$. Thus the method of descent implies that S_n is always true. One can show that the two principles are equivalent but we have traversed to far away from Euclid already.

Returning to the fundamental theorem of arithmetic we have shown that if A is not the unit then A can be written as $q_1 q_2 \cdots q_n$ with q_i a prime for $i = 1, ..., n$. Since the q_i are not necessarily distinct we can take $p_1, ..., p_r$ to be the distinct ones and group those together to get $A = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$ (here e_1 is the number of i such that $p_1 = q_i$, e_2 is the number of i such that $p_2 = q_i$,...). We are now ready to prove the uniqueness. The crux of the matter and is Proposition 30 of Book VII which says:

If two numbers by multiplying one another make some number, and any prime number measure the product, it will measure one of the original numbers.

Let us see how this proposition implies our assertion about uniqueness. We will prove it using the principle of mathematical induction. The assertion P_n is that if n is not one then up to order there is only one expression of the desired form. Notice that P_1 doesn't say anything so it is true (by default).

Suppose we have proved P_n for $1 \leq m < n$. Assume that $n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$ and $n = q_1^{f_1} q_2^{f_2} \cdots q_s^{f_s}$. with $p_1, ..., p_r$ distinct primes and $q_1, ..., q_s$ distinct primes. Then we must show that $r = s$ and we can reorder $q_1, ..., q_r$ so that $q_i = p_i$ and $f_i = e_i$ for all $i = 1, ..., r$. Since p_1 divides n we must have p_1 divides $q_1(q_1^{f_1-1} q_2^{f_2} \cdots q_s^{f_s})$. Thus p_1 divides q_1 or $q_1^{f_1-1} q_2^{f_2} \cdots q_s^{f_s}$ by Proposition 30. If it divides q_1 it is equal to q_1. Otherwise since it cannot divide $q_1^{f_1-1}$ it divides $q_2^{f_2} \cdots q_s^{f_s}$. Proceeding in this way we eventually see that there must be an index i so that $p_1 = q_i$. Relabel so that $i = 1$. Then we see that if $m = n/p_1$ then $m = p_1^{e_1-1} p_2^{e_2} \cdots p_r^{e_r}$, and $m = q_1^{f_1-1} q_2^{f_2} \cdots q_s^{f_s}$. If $m = 1$ then $n = p_1 = q_1$. Otherwise $1 < m < n$ so P_m is true. Hence $s = r$ and $f_1 - 1 = e_1 - 1$ and the other q_i can be rearranged to get the conclusion $q_i = p_i$ and $f_i = e_i$ for $i = 2, ..., r$.

So to complete the discussion of our Proposition we need only give a proof of Proposition 30 Book VII. This proposition rests on his theory of proportions (now rational numbers). We will give an argument which uses negative numbers (jumping at least 1500 years in our story). We will assume here that the reader is conversant with integers $(0, \pm 1, \pm 2, ...)$. Our argument is based on Propositions 1 and 2 Book VII given in the following form:

If x,y are numbers that are relatively prime (prime to each other) then there exist integers a, b such that $ax + by = 1$.

We follow the procedure in the argument that demonstrates Propositions 1 and 2 of Book VII . If $x > y$ then the first step is $x - y$. We assert that at each stage of this subtraction of the lesser from the greater we have a pair of numbers $ux+vy$ and $zx+wy$ with u, v, z, w integers. At step one this is clear. So suppose this is so at some step we show that it is so at the next step. So if $ux + vy$ and $zx + wy$ are what we have at some step then if (say) $ux + vy > zx + wy$ then at the next step we have $(ux + vy) - (zx + wy)$ and $zx + wy$. That is $(u - z)x + (v - w)y$ and $zx + wy$. According to Propositions 1 and 2 Book VII this will eventually yield 1.

We will now demonstrate Proposition 30 Book VII. Suppose that p is a prime, a, b are numbers and p divides ab, but p does not divide a. Then p and a are relatively prime. Thus there exist integers u, v so that $up + va = 1$. Now $b = upb + vab$ since and $ab = pc$ we see that $b = ubp + vcp = (ub + vc)p$.

We will also describe how Euclid proves Proposition 30. Let C be the product of A and B and assume that D is a prime dividing C then C is the product of D and E. Now assume that A and D are prime to each other (since D is prime this means that D does not measure A). Then D, A and B, E are in the same proportion. Since D is prime and A all pairs in the same proportion to D, A are given as multiples FD, FA (this is a combination of Propositions 20 and 21 in Book VII) thus D measures B.

1.3.3 Exercises.

1. Use the method of Propositions 1 and 2 of Book VII to calculate the "greatest common measure" of 315 and 240 and of 273 and 56..

2. Read the original proof of Proposition 30 Book VII. Explain how it differs from the argument given here. Also explain in what sense the two proofs are the same.

3. Use the principal of mathematical induction to show

(a) $1+4+9+....+n^2=\frac{n(n+1)(2n+1)}{6}$.

(b) $1+2+4+...+2^n=2^{n+1}-1$.

4. Use the material of this section to show that if $\frac{a}{b}$ is a fraction then it can be written uniquely in the form $\frac{c}{d}$ with c,d in lowest terms (relatively prime). In other words complete the discussion of the proof of Proposition 30 Book VII).

5. Assume that $1+2^m+...+n^m=p_m(n)$ with p_m a polynomial of degree $m+1$ in n. Set up a formula of the form of (a),(b) for the sum of cubes. Prove it by induction. Why do you think that the assertion about p_m is true?

6. Use the method of descent to prove that there is no rational number $\frac{a}{b}$ so that $\left(\frac{a}{b}\right)^2=2$. Hint: Let P_n be the statement that there is no m such that $n^2=2m^2$. Use Proposition 30 show that if $n^2=2m^2$ then n is even. Use this to show that if P_n is false then P_m is false for m such that $n^2=2m^2$.

1.4 Perfect numbers and primes.

1.4.1 The result in Euclid.

Perfect numbers are not a central topic in mathematics. However, their study has led to some important consequences. As we saw Euclid devoted one of his "precious" 22 definitions in Book VII to this concept. We recall that a perfect number is a number that has the property that the sum of its divisors (including 1 but not itself) is equal to itself. Thus 1 has as divisor 1 which is itself so it is not perfect. 2 has divisor 1 other than itself as does 3 and 5 so 2,3,5 are not perfect. Four has divisors 1,2 other than itself so it is not perfect. 6 has divisors 1,2,3 other than itself so it is perfect. Thus the smallest perfect number is 6. One can go on like this the next is 28 whose factors other than itself 1,2,4,7,14. It is still not known if there are only a finite number of perfect numbers. Euclid in Proposition 36 Book IX gave a "method" that generates perfect numbers. Let us quote the proposition.

> *If as many numbers as we please beginning from an unit be set out continuously in double proportion, until the sum becomes prime, and if the sum multiplied by the last make some number, the product will be perfect.*

This says that if $a=1+2+4+...+2^n$ is prime then 2^na is perfect. Notice that as Euclid gives the result it allows us to discover perfect numbers if we know that certain numbers are prime. We will now try it out. Euclid does not think of 1 as prime. $1+2=3$ is prime. $2\cdot3=6$ is thus perfect. $1+2+4=7$ is prime so $4 \cdot 7 = 28$ is perfect. $1+2+4+8 = 15$ not prime. $1+2+4+8+16 = 31$ is prime so $16 \cdot 31 = 496$ is perfect. We now check this because it tells us why the proposition is true. Write out the prime factorization of 496 (which we have seen is unique in the last section as $2^4 31$. Thus the divisors of 496 other than itself are $1, 2, 2^2 = 4, 2^3 = 8, 2^4 = 16, 31, 2 \cdot 31 = 62$, $2^2 \cdot 31 = 124$, $2^3 \cdot 31 = 248$. Add them up and we see that Euclid was correct.

The example of 496 almost tells us how to demonstrate this assertion of Euclid. If $a = 1+2+...+2^n$ is prime then the factors of $2^n a$ are $1, 2, ..., 2^n, a, 2a, ..., 2^{n-1}a$. So the sum of the factors is $1 + 2 + ... + 2^n + a + 2a + ... + 2^{n-1}a$. This is equal to $a + (1 + 2 + ... + 2^{n-1})a$. Now we observe that Exercise 2 (c) of section 1.4 implies that $1 + 2 + ... + 2^{n-1} = 2^n - 1$. Thus the sum of the factors is $a + (2^n - 1)a = a + 2^n a - a = 2^n a$.

1.4.2 Some examples.

This proposition is beautiful in its simplicity and we will see that the Swiss mathematician Leonhard Euler (1707-1783) proved that every even perfect number is deducible from this Proposition. The catch is that we have to know how to test whether a number is prime. We have noted that $1 + 2 + 4 + ... + 2^n = 2^{n+1} - 1$. Thus we are looking for numbers of the form $2^m - 1$ that are prime. Let us make an observation about this point. If $m = 2k$ were even then $2^m - 1 = 2^{2k} - 1 = (2^k + 1)(2^k - 1)$. If $k = 1$ then we have written $3 = 3 \cdot 1$ so if $m = 2$, $2^m - 1$ is prime. If $k > 1$ then $2^k - 1 > 1$ and $2^k + 1 > 1$ so the number is not prime. We therefore see that if $2^m - 1$ is prime and $m > 2$ then m must be odd.

To get 496 we used $2^5 - 1 = 31$. The next number to check is $2^7 - 1 = 127$. We now check whether it is prime. We note that if $a = bc$ and $b \leq c$ then $b^2 \leq a$. This is so because if $b \leq c$ then $b^2 \leq bc = a$. Thus we need only check whether 127 is divisible by 2,3,5,7,11 (since $12^2 = 144 > 127$). Since it is not we have another perfect number $127 \cdot 64 = 8128$. Our next candidate is $2^9 - 1 = 511 = 7 \cdot 73$.

We see that $2^2 - 1$, $2^3 - 1$, $2^5 - 1$, $2^7 - 1$ are prime but $2^9 - 1$ is not. One might guess from this that if $2^m - 1$ is prime then m must be prime. Obviously we are guessing on the basis of very little information. However, this is the way mathematics is actually done. So suppose that $m = ab$, $a > 1$, $b > 1$ we wish to see if we can show that $2^{ab} - 1$ is composite. Set $x = 2^a$ then our number is $x^b - 1$. We assert that $x^b - 1 = (x - 1)(1 + x + ... + x^{b-1})$. One way to do this is to remember long division of polynomials the other is to multiply out

$$(x - 1)(1 + x + ... + x^{b-1}) = x + x^2 + ... + x^b - 1 - x - ... - x^{b-1}.$$

Then notice that $x, x^2, ..., x^{b-1}$ subtract out and we have $x^b - 1$ left. Armed with this observation we can show the following proposition.

If $p = 2^m - 1$ is prime then m is prime.

If $m = ab$ and $a > 1$, $b > 1$ then setting $x = 2^a$ we see that $p = x^b - 1 = (x-1)(1+x+...+x^{b-1}) = cd$, $c = x - 1 > 1$ and $d = 1+x+x^2+...+x^{b-1} > 1$.

Our next candidate is 11. But $2^{11}-1=2047=23\cdot 89$. Using Mathematica (or any program that allows one to do high precision arithmetic) one can see that among the primes less than or equal to 61, 2^p-1 is prime for exactly $p=2,3,5,7,13,17,19,31,61$. Notice the last yields a prime $2^{61}-1=2305843009213693951$. We note that at this writing (2002) the largest known prime of the form 2^p-1 is $2^{13466917}-1$(Michael Cameron, 2001 with the help of *GIMPS* -Great Internet Mersenne Prime Search).

1.4.3 A theorem of Euler.

We give the theorem of Euler that shows that if a is a perfect even number then a is given by the method in Euclid. Write $a=2^m r$ with $r>1$ odd. Suppose that r is not prime. Let $1<a_1<...<a_s<r$ be the factors of r. Then the sum of the factors of a other than a is

$$(1+2+...+2^m)+(1+...+2^m)a_1+...+(1+...+2^m)a_s+(1+...2^{m-1})r$$

$$=2^{m+1}-1+(2^{m+1}-1)a_1+...+(2^{m+1}-1)a_s+(2^m-1)r.$$

Since we are assuming that a is perfect this expression is equal to a. Thus

$$(2^{m+1}-1)(1+a_1+...+a_s)+(2^m-1)r=2^m r.$$

We therefore have the equation

$$(2^{m+1}-1)(1+a_1+...+a_s)=r.$$

From this we conclude that $1+a_1+...+a_s$ is a factor of r other than r. But then $1+a_1+...+a_s\leq a_s$. This is ridiculous.. So the only option is that r is prime. Now we have

$$(2^{m+1}-1)+(2^m-1)r=2^m r.$$

So as before, $r=2^{m+1}-1$. This is the assertion of the proposition.

1.4.4 The Sieve of Eratosthenes.

In light of these results of Euler and Euclid, the search for even perfect numbers involves searching for primes p with 2^p-1 a prime. So how can we tabulate primes? The most obvious way is to make a table of numbers $2,...,n$ and check each of these numbers to see if it is divisible by an earlier number on the list. This soon becomes very unwieldy. However, we can simplify our problem by observing that we can cross off all even numbers, we can then cross off all numbers of the form $3\cdot n$ then $5\cdot n$ then $7\cdot n$, etc. This leads to the Sieve of Eratosthenes (230 BC)

1 2 3 4 5 6 7 8 9 10 11 12 13 ...

2 4 6 8 10 12 14 16 18 20 22 24 26 28 30...

3 6 9 12 15 18 21 24 27 30 33 36 39 42 45 48 51 54 57 60...

We cross out all numbers in the first row that are in the second or third row and have 1 5 7 11 13 17 19 23 25 29 31 35 37 41 43 47 49 53 55 59 61 ... The next sequence of numbers to check is the multiples of 5

5 10 15 20 25 30 35 40 45 50 55 60 65 ...

This reduces the first row to 1 7 11 13 17 19 23 29 31 37 41 43 47 49 53 59 61 ...

Now the number to check is 7

7 14 21 28 35 42 49 56 63 ...

Deleting these gives 1 11 13 17 19 23 29 31 37 41 43 53 59 61 ...

The next to check is thus 11

11 22 33 44 55 66 ...

We note that $11^2 = 121 > 61$. Thus we see that the primes less than or equal to 61 are 2,3,5,7,11,13,17,19,23,29,31,37,41,43,53,59,61.

The main modern use of the Sieve of Eratosthenes is as a benchmark to compare the speed of different digital computer systems. Most computational mathematics programs keep immense tables of primes and this allows them to factor relatively large numbers. For example to test that $2^{31} - 1$ is prime one notes that this number is of the order of magnitude of 2.4×10^9 thus we need only check whether it is divisible by primes less then or equal to about 5×10^4 so if our table went to 50,000, the test would be almost instantaneous. However, for $2^{61} - 1$ which is of the order of magnitude of 2.4×10^{19} the table would have to contain the primes less than or equal to about 50 billion. This is not reasonable for the foreseeable future. Thus other methods of testing primality are necessary. Certainly, computer algebra systems use other methods since, say Mathematica, can tell that $2^{61} - 1$ is a prime in a few seconds. Using Mathematica one can tell that $2^{89} - 1 = 618970019642690137449562111$ is a prime. This gives the next perfect number $2^{88}(2^{89} - 1)$ which is (base 10)

$$191561942608236107294793378084303638130997321548169216.$$

We will come back to the question of how to produce large primes and factoring large numbers later. In the next section we will give a method of testing if a number is a prime. We will see that the understanding of big primes has led to "practical" applications such as public key codes which today play an important role in protecting information that is transmitted over open computer networks.

1.4.5 Exercises.

1. Use the Sieve of Eratosthenes to list the primes less than or equal to 1000.

2. Write a program in your favorite language to store an array of the primes less than or equal to 500,000. Use this to check that $2^{31} - 1$ is prime.

3. The great mathematician Pierre Fermat (1601-1665) considered primes of the form $2^m + 1$. Show that if $2^m + 1$ is prime then $m = 2^k$. (Hint: If $m = ab$

with $a > 1$ and odd then set $x = 2^b$. $2^m+1 = x^a+1$. Now $-((-x)^a-1) = x^a+1$ since a is odd. Use the above material to show that $2^m + 1$ factors.) This gives $2^1 + 1 = 3$, $2^2 + 1 = 5$, $2^4 + 1 = 17$, $2^8 + 1 = 257$,... (so far so good). Fermat guessed that a number of the form $2^{2^m}+1$ is prime. Use a mathematics program (e.g. Mathematica) to show that Fermat was wrong.

4. The modern mathematician George Polya gave an argument for the proof that there are an infinite number of primes using the Fermat numbers $F_n = 2^{2^n} + 1$. We sketch the argument and leave the details as this exercise. He asserted that if $n \neq m$ then F_n and F_m are relatively prime. To see this he observes that

$$x^{2r} - 1 = (x^r - 1)(x^r + 1)$$

and so

$$x^{2^k} - 1 = (x^{2^{k-1}} - 1)(x^{2^{k-1}} + 1) = (x^{2^{k-2}} - 1)(x^{2^{k-2}} + 1)(x^{2^{k-1}} + 1)$$

$$= ... = (x^2 - 1)(x^2 + 1)(x^4 + 1) \cdots (x^{2^{k-1}} + 1) =$$

$$(x - 1)(x^1 + 1)(x^2 + 1)(x^4 + 1)(x^8 + 1) \cdots (x^{2^{k-1}} + 1).$$

If $n > m$ then $n = m + k$ so $2^n = 2^m 2^k$. Hence

$$F_n = (2^{2^m})^{2^k} + 1.$$

So (setting $x = 2^{2^m}$) $F_n - 2 = (x + 1)K$ with

$$K = (x^1 + 1)(x^2 + 1)(x^4 + 1)(x^8 + 1) \cdots (x^{2^{k-1}} + 1).$$

Thus $F_n - 2 = F_m K$. So if p is a prime dividing F_n and F_m then p must divide 2. But F_n is odd. So there are no common factors. Now each F_n must have at least one prime factor, p_n. We have $p_1, p_2, ..., p_n, ...$ all distinct.

5. We say that a number, n, is k perfect if the sum of *all* of its factors $(1, ..., n)$ is kn. Thus a perfect number is 2-perfect. There are 6 known 3-perfect numbers. Can you find one?

1.5 The Fermat Little Theorem.

In the last section we saw how the problem of determining perfect numbers leads almost immediately to the question of testing if a large number is a prime. The most obvious way of testing if a number a is prime is to look at the numbers b with $1 < b^2 \leq a$ and check if b divides a. If one is found then a is not a prime. It doesn't take much thought to see that this is a very time consuming method of a is really big. One modern method for testing if a is *not* a prime goes back to a theorem of Fermat. The following Theorem is known as the Fermat Little Theorem.

1.5.1 The theorem.

If p is a prime and if a is a number that is not divisible by p then $a^{p-1}-1$ is divisible by p.

Let us look at some examples of this theorem. If $p=2$ and a is not divisible by 2 then a is odd. Hence $a^{p-1}-1=a-1$ is even so divisible by 2. If $p=3$ and a is not divisible by p then $a=kp+1$ or $a=kp+2$ by the Euclidean algorithm. Thus $a^{p-1}-1$ is either of the form $(3k+1)^2-1$ or $(3k+2)^2-1$. In the first case if we square out we get $9k^2+6k+1-1=3(3k^2+2)$. In the second case we have $9k^2+12k+4-1=3(3k^2+4k+1)$. We have thus checked the theorem for the first 2 primes (2,3). Obviously, one cannot check the truth of this theorem by looking at the primes one at a time (we have seen that Euclid has demonstrated that there are an infinite number of primes). Thus to prove the theorem we must do something clever. That is demonstrate divisibility among a pair of numbers about which we are almost completely ignorant.

1.5.2 A proof.

We now give such an argument. If a is not divisible by p then ia is not divisible by p for $i=1,...,p-1$ (Euclid, Proposition 30 Book VII). Thus the Euclidean algorithm implies that if $1\leq i\leq p-1$ then $ia=d_ip+r_i$ with $1\leq r_i\leq p-1$. If $i>j$ and $r_i=r_j$ then $ia-ja=d_ip+r_i-d_jp-r_j=(d_i-d_j)p$. So $(i-j)a$ is divisible by p. Since we know that this is not true $(1\leq i-j\leq p-1)$ we conclude that if $i\neq j$ then $r_i\neq r_j$. This implies that $r_1,...,r_{p-1}$ is just a rearrangement of $1,...,p-1$.

Before we continue the proof let us give some examples of the rearrangements. We look at $a=2$, $p=3$. Then $a=0\cdot 3+2$, $2\cdot a=4=1\cdot 3+1$. Thus $r_1=2$, $r_2=1$. Next we look at $a=3$ and $p=5$. Then $3=0\cdot 5+3$, $6=1\cdot 5+1$, $9=1\cdot 5+4$, $12=2\cdot 5+2$. Thus $r_1=3$, $r_2=1$, $r_3=4$, $r_4=2$.

We can now complete the argument. Let us denote by s_j for $1\leq j\leq p-1$ numbers given by the rule that $r_{s_j}=j$. Thus in the case $a=2$, $p=3$, $s_1=2$, $s_2=1$. In the case $a=3$, $p=5$ we have $s_1=2$, $s_2=4$, $s_3=1$, $s_4=3$. Then we consider

$$a\cdot(2\cdot a)\cdot(3\cdot a)\cdots((p-1)\cdot a).$$

We can write this in two ways. One is

$$1\cdot 2\cdot\cdots(p-1)\cdot a^{p-1}.$$

The second is

$$(d_{s_1}p+1)\cdot(d_{s_2}p+2)\cdots(d_{s_{p-1}}p+(p-1)).$$

If we multiply this out we will get many terms but by inspection we can see that the product will be of the form

$$1\cdot 2\cdots(p-1)+c\cdot p.$$

We are getting close! This implies that

$$1 \cdot 2 \cdot \cdots (p-1) \cdot a^{p-1} = 1 \cdot 2 \cdots (p-1) + c \cdot p.$$

If we bring the term $1 \cdot 2 \cdots (p-1)$ to the left hand side and combine terms we have

$$1 \cdot 2 \cdots (p-1) \cdot (a^{p-1} - 1) = c \cdot p.$$

Thus p divides the left hand side. Since p can't divide any one of 1,2,...,$p-1$, we conclude that p divides $a^{p-1} - 1$.

1.5.3 The tests.

This leads to our test. If b is an odd number and $2^{b-1} - 1$ is not divisible b then b is not prime. If b is odd and $2^{b-1} - 1$ is divisible by b then we will call b a *pseudo prime to base* 2. It is certain that if b is odd and not a pseudo prime to base 2 then b is not a prime. The aspect that is amazing about this test is that if we show that b *does not divide* $2^{b-1} - 1$ then there must be a number c with $1 < c < b$ that divides b about which we are completely ignorant!

On the other hand this test might seem ridiculous.. We are interested in testing whether a number b is prime. So what we do is look at the (generally) very much bigger number $2^{b-1} - 1$ and see if b divides it or not. This seems weird until you think a bit. In principal to check that a number is prime we must look at all numbers $a > 1$ with $a^2 \leq b$ and check whether they divide b. Our pseudo prime test involves long division of two numbers that we *already know*. That is the good news. The bad news is that the smallest pseudo prime to base 2 that is not a prime is $341 = 11 \cdot 31$ and it can be shown that if b is a pseudo prime to base 2 then so is $2^b - 1$. Thus there are an infinite number of pseudo primes to base 2. For example if p is prime then $2^p - 1$ is also a pseudo prime to base 2 (see Exercise 4 below).

Note that we could add to the test as follows. If b is odd and $2^{b-1} - 1$ is divisible by b we only know that b is a pseudo prime. We could then check whether 3 divides b and if it does we would know it is not a prime. If it doesn't we could check whether b divides $3^{b-1} - 1$. The smallest number that is not a prime but passes both tests is $1105 = 5 \cdot 13 \cdot 17$. One can then do the same thing with 5. We note that if we do this test for $2, 3, 5$ the non-primes less than $10,000$ that pass the test are $\{1729, 2821, 6601, 8911\}$.

This leads to a refined test that was first suggested by Miller-Rabin. Choose at random a number a between 1 and $b - 1$. If the greatest common divisor of a and b is not one then b is not prime. If a and b are relatively prime but $a^{b-1} - 1$ is not divisible by b then b is not prime. If one repeats this k times and the test for being composite fails then the probability of b being composite is less than or equal to $\frac{1}{2^k}$. Thus if k is 20 the probability is less than one in a million. Obviously if we check all elements a less than b then we can forget about the Fermat part of the test. The point is that the number b is *very* big and if we do 40 of these tests we have a probability of better than 1 in 10^{12} that we have a prime.

A number, p, that is not a prime but satisfies the conclusion of Fermat's theorem for all choices of a that are relatively prime to p is called a Charmichael number the smallest such is 561. Notice that $561 = 3 \cdot 11.17$.

One further sharper test(the probabilities go to 0 faster and have strictly less failures than the Miller-Rabin test) is the Solovay-Strassen probabilistic test. We can base it on the proof we gave of Fermat's Little Theorem. Suppose that a and b are relatively prime and b is odd and bigger than 1. For each $1 \leq j < b$ we write

$$ja = m_j b + r_j$$

with $0 \leq r_j < b$. We note that r_j can't be zero since then b divides ja. Since b has no prime factures in common with a this implies that b divides j. This is not possible since $0 < j < b$. Thus, as before the numbers $r_1, ..., r_{b-1}$ form a reordering of $1, 2, ..., b-1$. We denote by π the number that is gotten by multiplying together the numbers $r_j - r_i$ with $j > i$ and $j < b$. Then since $r_1, ..., r_{b-1}$ is just a rearrangement of $1, ..., b-1$ we see that π is just ±1 times the number we would get without a rearrangement. We write $J(a, b)$ for 1 if the products are the same and -1 if not. We now consider the product $ja - ia)$ for $j > i$ and $1 \leq j < b$ then as we argued above we see that this number is

$$\pi + cb$$

with c a number. This says that if Δ is the product of $j - i$ over the same range then

$$\Delta a^{\frac{(b-1)(b-2)}{2}} = J(a, b)\Delta + cb.$$

This implies that if b is prime that $a^{\frac{(b-1)(b-2)}{2}} - J(a, b)$ is divisible by b. Now $n - 2$ is odd so $J(a, b) = J(a, b)^{n-2}$. Hence we see that if b is prime then

$$a^{\frac{b-1}{2}} - J(a, b) = db$$

for some number d. This leads to the test. We say that a number a between 2 and $b - 1$ is a witness that b is not prime if a and d are not relatively prime or $a^{\frac{b-1}{2}} - J(a, b)$ is not divisible by b. One can show that if there are no witnesses then b is prime. One can also prove that if b is not prime then more than half of the numbers a between 2 and $b - 1$ are witnesses. The test is choose a number a between 2 and $b - 1$ at random. If a is a not a witness that b is not prime then the probability is strictly less than $\frac{1}{2}$ that b is composite. Repeating the test say 100 times and not finding a witness will allow us to believe with high probability that b is prime.

The point of these statistical tests is that if we define $\log_2(n)$ to be the number of digits of n in base 2 then the prime number theorem (J. Hadamard and de Vallée Poussin 1896–we will talk about this later) implies that if N is a large number then there is with high probability a prime between N and $\log_2(N)$. For example, $N = 56475747478568$ then $\log_2(N) = 45$ and 56475747478601 is a prime. Thus to search for a prime with high probability with say 256 digits base 2 choose one such number (at "random"), N, then use the statistical tests on the numbers between N and $N + 256$.

The reader who has managed to go through all of this might complain that the amount of calculation indicated in these tests is immense. When we talk about modular arithmetic we will see that this is not so. In fact these tests can be implements very rapidly. As a preview we consider the amount of calculation to test that a number b is a is a 2-pseudo prime. We calculate 2^{b-1} as follows: We write out $b-1$ in base 2 say $b-1 = c_1 2 + c_2 4 + ... + c_n 2^n$ with c_i either 0 or 1. We then

$$2^{b-1} = (2^2)^{c_1}(2^4)^{c_2}\cdots(2^{2^2})^{c_n}.$$

As we compute the products indicated we note that if m is one of the intermediate products and if we apply division with remainder we have

$$m = ub + r$$

with $0 \leq r < b$. In the test we can ignore multiples of b. Also we use the fact that $2^{2^{m+1}} = (2^{2^m})^2$. And the 2^{2^m} can be replaced by its remainder after division by b. Let r_m be the remainder for 2^{2^m}. Thus if we have multiplied the first k terms and reduced to a number less than b using division with remainder to have the number say s if $c_{k+1} = 1$ we multiply s by r_k and then take the remainder after division by b. We therefore see that there are at most n operations of division with remainder by b and never multiply numbers as big as b. We will see that a computer can do such a calculation very fast even if b has say 200 binary digits. We give an example of this kind of calculation consider the number $n = 65878161$. Then 2^{n-1} is an immense number but if we follow the method described we have the binary digits of $n-1$ (written with the powers of 2 in increasing order) are

$$\{0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1\}.$$

The method says that each time we multiply we take only the remainder after division by n. We thereby get for the powers

$$\{/2, 4, 16, 256, 65536, 12886831, 1746169, 1372837, 38998681, 33519007, 56142118,$$
$$28510813, 45544273, 49636387, 27234547, 48428395, 5425393, 65722522,$$
$$46213234, 3252220, 64423528, 16511530, 46189534, 45356743, 15046267, 47993272\}.$$

Now the intermediate products are (we include the terms where the digit is 0)

$$\{1, 1, 1, 65536, 65536, 65536, 46555867, 46555867, 46555867, 46555867,$$
$$18206659, 42503458, 24542662, 24542662, 24542662, 54699517, 54699517,$$
$$29732113, 728509, 728509, 38913619, 36121177, 1794964, 30837667, 23401021\}.$$

The last number is not one so the number n is not a prime. This seems like a lot of computation but most modern personal computers can do these calculations instantly. It turns out the $n = 7919 \times 8319$. So finding a factor by trial and error would have involved more computations. We also observe that the same method can be used for any choice of a using $a^{2^{m+1}} = (a^{2^m})^2$.

1.5.4 Exercises.

1. Make a large list of pseudo primes base 2 less than or equal to 1000. Compare this with a list of primes less than or equal to 1000. (You will want to use a computer for this.)

2. If n is any positive integer show that there exists a consecutive list of composite integers of length n. (Hint: If we set $(n+1)! = (n+1)n(n-1)\cdots 2$ then $(n+1)!+2, (n+1)!+3, ..., (n+1)!+n+1$ are all composite.) For each $n = 2, 3, 4, 5, 6, 7, 8, 9$ find the consecutive list of primes that starts with the smallest number (for example if $n = 3$ the answer is $8, 9, 10$). Why do we need to only check n odd?

3. Calculate the rearrangement of 1,2,...,6 that corresponds to $a = 2$ and $p = 7$ as in the proof of the Little theorem. Use this to calculate $J(2, 7)$.

4. Given a and p as in Fermat's Little theorem and $r_1, ..., r_{p-1}$ and $s_1, ..., s_{p-1}$ show that if $1 < a < p$ then $r_1 = a$ and $s_1 \cdot r_1 = u \cdot p + 1$ with u a whole number.

5. Show that if p is a pseudo prime to base 2 then so is $2^p - 1$. (Hint: If $q = 2^p - 1$ then

$$2^{q-1} - 1 = 2^{2^p-2} - 1 = 2^{2(2^{p-1}-1)} - 1.$$

Now p divides $2^{p-1} - 1$. So $2^{p-1} - 1 = cp$. Thus

$$2^{q-1} - 1 = 2^{2cp} - 1 = x^{2c} - 1$$

with $x = 2^p$.)

6. We note that $1 + 2^2 + 1 = 6$ (so divisible by 3), $1 + 2^4 + 3^4 + 4^4 + 1 = 355$ (so divisible by 5). Show more generally that if p is prime then

$$1^{p-1} + 2^{p-1} + ... + (p-1)^{p-1} + 1$$

is divisible by p. It has been shown that if p satisfies this condition (that it divides the above sum) then it has been shown by Giuga(Giuga, G. "Su una presumibile propertietà caratteristica dei numeri primi." Ist. Lombardo Sci. Lett. Rend. A 83, 511-528, 1950) that p is a Charmichael number. He also conjectured that such a number must, in fact be prime. This has been checked for $p < 10^{13800}$ (Borwein, D.; Borwein, J. M.; Borwein, P. B.; and Girgensohn, R. "Giuga's Conjecture on Primality." Amer. Math. Monthly 103, 40-50, 1996).

7. Use a package like Mathematica or Maple to show that 341 is a pseudo prime to base 2 and that $2^{1104} - 1$ and $3^{1104} - 1$ are both divisible by 1105.

8. To do this problem you should use a computer mathematics system. Calculate the remainder of dividing 2^{n-1} by n for $n =$
579833790077893015263432471098694218875498494876858922371038810170007677183040183965105133072849587678042834295677745661721093871. Use the outgrowth of the calculation to deduce that n is not a prime.

1.6 Large primes and cryptography.

In the last section we saw that large primes appear naturally in the "unnatural" problem of finding perfect numbers. Large primes have also become an important part of secure transmission of data. Most modern cryptographic systems involve two “keys” one to be used to encode and the other to decode messages. Public key systems have a novel aspect in that the information necessary to encode a message is in principle known to everyone. But the information to decode the message is only known to the person the intended recipient of the message. In other words, even if you know how to encode a message you still do not know how to decode a different message encoded by that method. Alternatively, even if you find a method of deciphering one message deciphering another is not easier. This is a seeming contradiction and although most believe that the methods now in use have this contradictory property there is no mathematical proof that this is so. This type of cryptography was first described by W.Diffie and M.E.Helman “New directions in cryptography,” IEEE Transactions in Information Theory IT-22 (1976),644-654.

One of the first “practical” implementations was due to Rivest, Shamir and Adelman (1978) and is called RSA. It is based on the hypothesis that the factoring of large numbers is much harder than multiplying large numbers. We will discuss this point and describe the implementation of RSA later in this section.

1.6.1 A problem equivalent to factorization.

In the RSA system a person (usually called Alice) chooses (or is assigned) two very large primes p and q. Alice calculates $n = pq$ and makes n public. She also chooses a number e (for encode) that has greatest common divisor 1 with the number $m = (p-1)(q-1)$ and such that $1 < e < m$. This number is also made public. The rest of the system involves enciphering messages using these two numbers (n, e). The point of the methods of enciphering is that to decode the message one must know a number $1 < d < m$ (for decode) such that $ed = rm+1$ for some integer r (note that the form of Proposition 1 Book VII in Euclid tells us that d exists. It is hypothesized that one cannot find d without knowing m. There are also probabilistic arguments that indicate that with high probability if we know d then we know m. The main point is thus the following Proposition:

If we know the number m then it is easy to factor n.

Before we demonstrate this we will interpret the line of thought. This assertion then says that with a high probability, deciphering the RSA cipher is at the same level of difficulty as factoring n. Since we have hypothesized that this is impractically hard we have implemented a public key system.

As for the Proposition, if we know m then we know $(p-1)(q-1) = pq-p-q+1$. Since we know $n = pq$ we therefore know $p+q$. Now, $(p+q)^2 - 2pq = (p-q)^2$ we see that we know $(p-q)^2$ at the same level of difficulty as squaring (which the ancient Egyptians thought was relatively easy) that we have hypothesized is much easier than factoring. The last step is to see that there is an “easy” method of recovering a if we know a^2. We will see that this is so below. Thus with little difficulty we have calculated $p + q$ and $p - q$. We can recover p and q by adding and subtracting and dividing by 2.

1.6.2 What do we mean by “hard” and “easy”?

Before we describe an implementation of RSA we will give a working explanation of the terms hard and easy. In what follows we will use the notation $\log_2(n)$ to mean the smallest k such that 2^k is greater than or equal to n. In other words $\log_2(n)$ is the number of operations necessary to write the number n in base 2. We will say that a procedure depending on integers $N_1, ..., N_d$ is *easy* if the there is a method for implementation (an algorithm) that takes a time to complete that is proportional to a fixed power (depending on the procedure) of $(\log_2(N_1) + ... + \log_2(N_d))$. If an operation is not easy then we say that it is *hard*. The study of hard and easy belongs to complexity theory. It is a formalism that is useful for testing whether good computational methods exist (or don’t exist). We will just touch the surface.

As our first example we consider the problem of comparing two numbers M and N. We assert that this takes at most $4(\log_2(N) + \log_2(M))$ operations. We will go through most of the (gruesome details for this case since it is the simplest. The reader should have *patience*). Indeed, it takes $\log_2(N)+\log_2(M)$ operations to write the two numbers. Once we have done this we know $\log_2(N)$ and $\log_2(M)$. We prove by induction on $r = \log_2(N) + \log_2(M)$ that it now takes at most $4(\log_2(M) + \log_2(N))$ operations to test whether N is bigger than M is smaller than N or is equal to N. If $r \leq 1$ then all we must do is look at the two indicated numbers which are 0 or 1. Assume for $r \leq s$ (the induction hypothesis). We now show that it is true for s. We first check that if $\log_2(N) > \log_2(M)$ (or $\log_2(N) < \log_2(M)$) then $N > M$ or $(N < M)$. This by the induction hypothesis we need at most $4(\log_2(\log_2(N)) + \log_2(\log_2(M))$ steps to check this. If we have strict comparison of the logs we are done in $2(\log_2(\log_2(N))+\log_2(\log_2(M))$ steps. Otherwise we now know that $\log_2(N) = \log_2(M)$ we now check the digits one by one from the top and look for the first place with one of M or N having a 1 and the other a 0 the one with the 1 is the larger. If we do the full number of steps we have equality. Thus we have done the comparison in at most $(\log_2(N)+\log_2(M))$ additional steps. Now we observe that $\log_2(n) \leq \frac{n}{2}$. If $n \geq 2$. If $n = 2$ this says that $1 \leq 1$. If it is true for n and if $n \neq 2^k - 1$ then $\log_2(n+1) = \log_2(n) \leq \frac{n}{2} < \frac{n+1}{2}$. Otherwise, $n = 2^k - 1$. So $\log_2(n+1) = k+1$. We are left with observing that $2^k \geq k+1$, for $k = 1, 2, ...$ For $k = 1$ we have equality. If $2^k \geq k+1$ then

$$2^{k+1} = 2(2^k) \geq 2(k+1) = 2k+2 \geq k+2.$$

This implies that

$$2(\log_2(\log_2(N)) + \log_2(\log_2(M)) \leq \log_2(N)) + \log_2(M)$$

So the total number of steps is at most

$$(\log_2(N) + \log_2(M)) + 2(\log_2(N)) + \log_2(M)) + (\log_2(N) + \log_2(N))$$

the first term for writing the two numbers, the second for comparing the number of digits and the third for the main comparison. Thus comparison in *easy* (as we should guess).

We now look at addition. We have numbers a and b write out the numbers in base 2 assume that a is the larger and fill out the digits of b by 0 (easy). This involves $2\log_2(a)$ operations. We write $n = \log_2(a)$. Now add the lowest digits, if one is 0, then put the other digit in the lowest position of the answer otherwise both are 1 so put a 0 in the lowest position and then look at the next digit of a if it is 0 change it to 1 if it is 1 change it to 0 then do the same operation on the next digit continue until you get to a 0 digit of a or to the top one which must be 1 and we would change it to 0 and add one more digit to a. This happens only if all the digits of a are 1 in this case $a = 2^n - 1$. So to add a and b you need only change the lowest digit of b to 0 and then add 2^n which involves at most 3 steps. This implies that we are either done in 3 steps or we need at most n operations to add the lowest digits. We then go to the next digit. We see that if we are adding at the rth digit we will need to at most the larger of 3 and $n - r$ easy operations. Thus the number of operations is at most $n + (n-1) + ... + 1 = \frac{n(n+1)}{2}$ easy operations. So addition is easy.

The next case is that division with remainder is easy. To see this we look at M and N and we wish to divide M into N. Comparison is easy. So if $M > N$ the division yields 0 with remainder N. If $M = N$ we get division 1 and remainder 0. Thus we nay assume $M < N$. Let m be the number of digits of M and n that of N. If $n = m$ then the division is 1 with remainder $N - M$ (subtraction is easy, you will do this in an exercise). Thus we can assume that $n > m$. Now multiply M by 2^{n-m} and (this just means putting $n - m$ zeros at the end of the base two expansion of m) subtract this from N. Getting N_1 with less than n digits. If $N_1 \leq M$ we are done otherwise do the operation again. After at most n of these steps we are done. Thus we must do at most n easy operations. So division with remainder is easy. We also note that similar considerations imply that addition, subtraction and multiplication are easy.

Consider Euclidian method of calculating the greatest common divisor (g.c.d.)of two numbers $n > m > 1$. first subtract m from n repeatedly until one has m or one has a number that is less than m. If the number is m then the g.c.d. is m. If not put $n_1 = m$ and m_1 equal to the number we have gotten and repeat. If $m_1 = 1$ then we know that the g.c.d. is 1. Thus The initial step involves about n/m subtractions. It also involves one division with remainder. If n is not divisible by m then m_1 is the remainder after division. Thus, if we use division rather than subtraction each step involves one division with remainder. Since each step reduces the bigger number to a number less than or equal to one half its size we see that the number of such operations is at most $\log_2(n)$. Thus it takes no more than $\log_2(n)$ times the amount of time necessary to calculate the division with remainder of n by m. By a hard operation on n or on $n > m$ we will mean an operation that involves more than a multiple of $\log_2(n)^k$ steps for each $k = 1, 2, 3, ...$ (the multiple could depend on k). Thus calculating the g.c.d. is easy.

To complete the line of reasoning in the previous subsection we show that if a is a positive integer then the calculation of the positive integer b such that $b^2 \leq a < (b+1)^2$ is easy b is called the *integer square root* of a.. The idea is to write out a to base 2. If the number of digits is 4 or less look it up in a table. If the number of digits is n which is odd $n = 2k+1$ then take as the first approximation to b the number 2^k if this satisfies the upper inequality we are done otherwise try $2^k + 2^{k-1}$ if it satisfies both inequalities we are done otherwise if it doesn't satisfy the lower one replace by $2^k + 2^{k-2}$ and continue the same testing to see if we leave the bit "on" or not. The involves calculating $2n$ squares so since $n - 1 = \log_2(a)$ and we have decided that squaring is easy we have shown that in this case calculating b is easy. If $n = 2k$ is even then look at the first 2 bits of a (the coefficients of the highest and next highest power of 2) then start with 2^{k-1} and use the same procedure.

Is anything hard? The implementation of RSA assumes that factoring a large number is hard. There is no proof of this assertion, but the best known methods of factorization take the order of magnitude of

$$2^{C(\log_2(N))^{\frac{1}{3}}}$$

steps.

1.6.3 An implementation of RSA.

Suppose that you are shopping on the internet and you must transmit your credit card number, C, to the merchant. You know that it is possible that "Joe Hacker" is watching for exactly this sort of transaction. Obviously, you would like to transmit the number in such a way that only the merchant can read it. Here is an RSA type method that might accomplish this task. The merchant chooses two big primes p and q (so big that they are both bigger than any credit card number) then forms the numbers $n = pq$ and $m = (p-1)(q-1)$. He also chooses e randomly between 1 and m that has greatest common divisor. 1 with m. He transmits the numbers n and e to your computer (and probably Joe's computer). Your computer then calculates the remainder that is gotten when C^e is divided by n. Call this number S. Your computer sends S to the merchant. This is what Joe sees. The merchant calculates the number d that has the property that $de = 1+mk$ for some k. He then calculates the remainder after division by n of S^d and has C we will explain this in the next paragraph. If Joe can calculate d then he also knows C. However, if the primes are very large we have seen that this is very improbable.

We now explain why $S^d = C + nh$ for some h. Neither p nor q divides C since it is too small. By definition of S,

$$C^e = S + ng$$

for some g. Thus $S = C^e - ng$. We therefore have

$$S^d - C^{de} = (C^e - ng)^d - C^{de}.$$

One can check the formula

$$x^d - y^d = (x-y)(x^{d-1} + x^{d-2}y + ... + xy^{d-2} + y^{d-1}).$$

by direct multiplication.

$$\begin{aligned}
& (x-y)(x^{d-1}+x^{d-2}y+...+xy^{d-2}+y^{d-1}) \\
= & \; x(x^{d-1}+x^{d-2}y+...+xy^{d-2}+y^{d-1})-y(x^{d-1}+x^{d-2}y+...+xy^{d-2}+y^{d-1}) \\
= & \; x^d+x^{d-1}y+...+x^2y^{d-2}+xy^{d-1} \\
& \quad -x^{d-1}y-...-x^2y^{d-2}-xy^{d-1}-y^d \\
= & \; x^d-y^d.
\end{aligned}$$

If we make the replacement $x = C^e - ng$ and $y = C^e$ in this formula we find that $S^d - C^{de}$ is divisible a multiple of $x - y = -ng$ and is thus divisible by n. Thus the remainder after dividing by n of S^d and C^{de} is the same. We note that

$$(C^e)^d = C^{de} = C^{1+mk} = C(C^m)^k.$$

Now $m = (p-1)(q-1)$ and $(C^{k(q-1)})^{p-1} = 1+ap$ by the Fermat Little Theorem. Similarly, $C^{mk} = 1 + bq$. Thus $C^{mk} - 1$ is divisible by both p and q hence by n. (See Exercise 2 below.) Thus $S^d = C(1 + cn) = C + un$ for some whole number u.

We will now do an example of this but with smaller numbers than those that would be in a practical implementation.. We take $p = 71$ and $q = 97$. Then $n = 6887$ and $m = 6720$. Choose $e = 533$. Then the "decoder" is $d = 2132$. If $C = 45$ then the remainder after division by n of C^e is 116. We note that 116^d has remainder 45 after division by n.

1.6.4 Fermat factorization.

RSA is based on the assumption that factoring big numbers is hard. How would we go about doing a factorization of a big number. If we knew that the number came from RSA we would then know that it has only two prime factors. Does this make the problem easier? Fortunately for the internet this doesn't seem to be the case. We will, however, look at a pretty good method of factoring now.

Suppose that n is an odd number and that $n = ab$ with $1 < a < b$. Set $t = \frac{a+b}{2}$ and $s = \frac{b-a}{2}$. Note that a and b are odd so t and s are whole numbers. We have

$$t^2 - s^2 = ab.$$

The reverse is also true, that is, if $1 \le s \le t$ and if $n = t^2 - s^2$ then if $a = t - s$ and $b = t + s$ then $n = ab$. This leads to a method. Start with the number n let g be its integer square root. if $g^2 = n$ we have factored the number into two smaller factors. Otherwise try $t = g + 1$ and calculate $t^2 - n$ if this number is a perfect square s^2 then apply the above observation Otherwise replace t by $t+1$ and try again. Keep this up until $t^2-n = s^2$. This is practical only if n has two factors that are very close together. This tells us that for the sake of security of RSA one must choose p and q far apart.

We will try this factorization out for the example we used above $n = 6887$ then the integral square root is 82. $82^2 = 6724$. $83^2-n = 2, 84^2-n = 169 = 13^2$.

So taking $a = 84 - 13 = 71$ and $b = 84 + 13 = 97$ we've found our original $p = a, q = b$.

There are many variants of this method that involve significant improvements in the number of operations necessary to do the factorization. However, the best known methods are hard in the sense of this section. In the next section we will show how a "change in the rules" allows for an "easy" factorization algorithm.

1.6.5 More approaches to factorization.

In 1994, Peter Shor published a proof that if a computer that obeys the rules of quantum mechanics could be built then it would be possible to factor large numbers easily. The subject of quantum computing would take us too far afield. However, one of the ingredients of Shor's approach can be explained here. We start with a large number, N. Choose a number y randomly. We calculate the remainder of division by N of y^x for $x = 0, 1, 2, ...$ and call that number $f(x)$. Then there is a minimal number $1 < T < N$ such that $f(x + T) = f(x)$ for all x. We call T the smallest period. If T is even we assert that $y^{\frac{T}{2}} + 1$ and N have a common factor larger than 1. We can thus use the Euclidean algorithm (which is easy in our sense above) to find a factor of N. Before we demonstrate that this works consider $N = 30$ and $y = 11$. Then $f(0) = 1, f(1) = 11$, $11^2 = 121 = 1 + 4 \cdot 30$, so $f(2) = 1 = f(0)$. Thus $T = 2$. Now $11^1 + 1 = 12$. The greatest common divisor of 12 and 30 is 6.

We will next check that this assertion about y, T, N is correct. We first note that

$$(y^{\frac{T}{2}} + 1)^2 = y^T + 2y^{\frac{T}{2}} + 1.$$

But $y^T = 1 + m \cdot N$ by the definition of T. Thus after division by N one gets the same remainder for $(y^{\frac{T}{2}} + 1)^2$ and for $2(y^{\frac{T}{2}} + 1)$. This implies that

$$(y^{\frac{T}{2}} + 1)^2 - 2(y^{\frac{T}{2}} + 1)$$

is evenly divisible by N. Thus so is

$$\left((y^{\frac{T}{2}} + 1) - 2\right)\left(y^{\frac{T}{2}} + 1\right) = \left((y^{\frac{T}{2}} - 1\right)\left(y^{\frac{T}{2}} + 1\right).$$

Thus if $y^{\frac{T}{2}} + 1$ and N have no common factor then $y^{\frac{T}{2}} - 1$ is evenly divisible by N. This would imply that $\frac{T}{2}$ which is smaller than T satisfies

$$f(x + \frac{T}{2}) = f(x).$$

This contradicts the choice of T as the minimal period.

There are several problems with this method. The most obvious is what happens if the minimal period is odd? It can be shown that the probability is small that one would make many consecutive choices of y with odd period. Thus the "method" is probabilistic. However, if you could decode RSA with probability, say, .6 , then you would be able to decode about 60% of the secure internet commerce. There, however, is a much more serious problem. There is no easy algorithm for computation of such periods. The standard ways of finding the T above are as difficult as the factoring algorithms. This is where quantum computing comes in. Shor's contribution was to assume that his computer allowed for "superpositions" (be patient we will know what this means later. For now if you don't know what this means read quantum mechanical operations.) of digits and that these superpositions obeyed the rules of quantum mechanics. Under these assumptions he proved that he could find the period easily.

1.6.6 Exercises.

1. Why are subtraction, multiplication and division with remainder easy (in the sense above)?

2. Show that if p, q are distinct primes that if p and q divide a then pq divides a.

3. Use Fermat factorization to factor each of the following numbers into a product of two factors 3819, 8051, 11921.

4. Suppose that you have intercepted a message that has been encoded in the following variant of RSA. Each letter in the message is translated into a number between 1 and 26. We will ignore case and all punctuation but spaces and a space is assigned 27. So a and A become 1 , z and Z become 26. Thus we would write NUMBER as $14, 21, 13, 2, 5, 18$. We think of this as a number base in base 28. (Here this number is $14{+}21{*}28{+}13{*}28^2{+}2{*}28^3{+}5{*}28^4{+}18{*}28^5 = 312\,914\,602$. We expand the number and write it to base 60. getting $22, 43, 40, 8, 24$. We then encode each digit using RSA with $n = 8051$ and $e = 1979$. This gives $269, 294, 7640, 652, 198$. Suppose that you know that $402, 2832$ was coded in this way. What did the original message say? (Even for relatively small numbers such as these you will almost certainly need a computer algebra package to do the arithmetic.)

5. A form of RSA is the standard method of sending secure information on the internet. Do you agree that it is secure?

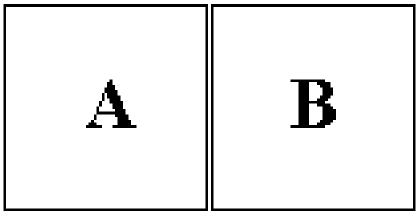
6. Consider all y between 10 and 20 and $N = 30$. Calculate the periods, T in the sense of the Shor algorithm (see the previous section).

2 The concept of geometry.

2.1 Early geometry.

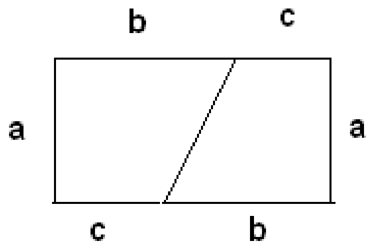
2.1.1 Babylonian areas.

In section 1.3 we alluded to the fact that Euclid did not look upon arithmetic as an outgrowth of simple counting. He rather looked upon it as arising from measurement of intervals with respect to a unit. The word geometry when analyzed has two parts geo for earth and metry for measurement. The earliest known record of geometry can be found in Babylonian tablets dated from about 3000 B.C. These tablets are concerned with calculating areas. One starts (as did Euclid) by measuring intervals with respect to a unit interval. The subject of these tablets was the calculation of areas bounded by four straight lines. If we think a bit about this question and decide that a square with side given by the chosen unit has unit area and if we take two of them on put then side by side (ore one on top of the other) then we have a rectangle with sides 2 and 1. It is reasonable to think that this rectangle has area 2.



Similarly we can put six such unit squares together and make a rectangle of sides 2 and 3 which has area 6. Thus if we have a rectangle of sides a and b then the area should be $a \cdot b$ (square units).

Obviously, not every area is as regular as a rectangle and the Babylonians concerned themselves with four sided figures that could be determined by 2,3 or 4 measurements.. Thus a rectangle of sides a and b is determined by two measurements. What about 3 measurements? Here imagine a rectangle of sides a, b and on one of the sides b an distance c from the side of length a is marked. One then joins the marked point with the endpoint of the other side of length b. One now has a figure that is sometimes called a *rectangular trapezoid.* Let us deduce the corresponding area.



The figure has sides labeled by a, b, c and the diagonal if we fold it over as in the picture above the two trapezoids fit together to make a rectangle of sides a and $b + c$. Thus the trapezoid is half of that rectangle and so we have shown that its area is $\frac{1}{2}(b + c)a$. This is the Babylonian formula.

It is still a subject of debate as to what the Babylonians meant by a figure determined by four measurements.. However, what seems to be agreed is that the formula that was used for the area does not jibe with any general notion of "four measures" since if the measurements are a, b, c, d then the formula they give is $\frac{(a+c)(b+d)}{4}$. This seems to be what they thought was the area of a general four sided figure with sides of lengths a, b, c, and d.

2.1.2 Right triangles.

As we saw the Babylonians understood Pythagorean triples. They in fact seemed to be aware of what we call the Pythagorean Theorem. In 1916 the German historian of mathematics Ernst Weidner translated a tablet from 2000 BC that contained the assertion that if a right triangle has legs a and b then the other side has length

$$c = a + \frac{b^2}{2a}.$$

This is not correct, in general, however we should recall that the Babylonians used the approximation

$$\sqrt{u^2 + v} = u + \frac{v}{2u}.$$

If we apply this formula we find that they are using

$$c = \sqrt{a^2 + b^2}.$$

2.1.3 Some Egyptian Geometry.

In the Moscow Papyrus (approximately 1700 BC) there is the following problem

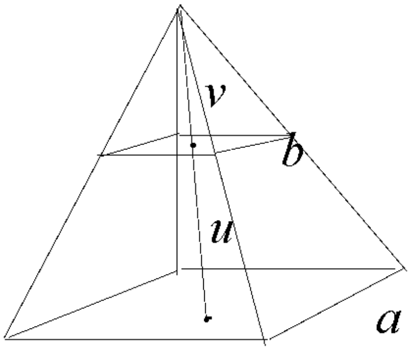
The area of a rectangle is 12 *, and the width is three quarters of the length, what are the dimensions?*

The solution was given in the following way. If we attach a rectangle of side one third of the smaller to the longer side to make the figure into a square then the area of the square would 16. Thus the longer side must be 4 and the shorter 3.

This method is what we now call "completing the square". This example indicates that the Egyptians understood rectilinear areas. However in the same Papyrus there is the following:

If you are told: A truncated pyramid of 6 *for the vertical height by* 4 *on the base by* 2 *on the top. You are to square this* 4*, result* 16*. You are to double* 4 *result* 8*. You are to square* 2 *result* 4*. You are to add the* 16*, the* 8*, and the* 4*, result* 28*. You are to take one third of* 6*, result* 2*. You are to take* 28 *twice result* 56*. You will find it right.*

Here the scribe clearly has in mind a truncated pyramid of the type that we know that the Egyptians built. That is the base is square and the top is parallel to the base and centered over it.. If we write u for the height (6) and a for the side of the base (4) and b for the side of the top (2). Then the scribe has written: $\left(b^2+ab+a^2\right)\left(\frac{u}{3}\right)$. Which is the correct formula. We will just indicate how it follows from the formula for the volume of a pyramid of base a and height h, $\frac{ha^2}{3}$. Consider the picture below



Then the total volume is $\frac{(u+v)a^2}{3}$. Now the theory of similar triangles (see Thales fourth Theorem below or Proposition 4 Book VI in Euclid) we have

$$\frac{u+v}{a}=\frac{v}{b}.$$

The desired volume is thus

$$\frac{(u+v)a^2}{3}-\frac{vb^2}{3}=\frac{va^3}{3b}-\frac{vb^3}{3b}.$$

We now rewrite the identity just used as

$$\frac{u}{a}+\frac{v}{a}=\frac{v}{b}$$

that is

$$\frac{u}{a} = v\left(\frac{1}{b} - \frac{1}{a}\right) = \frac{(a-b)v}{ab}$$

this gives

$$v = \frac{b}{a-b}.$$

Substituting this into the formula we have for the desired area yields

$$u\frac{a^3 - b^3}{3(a-b)}.$$

We now note that this implies the Egyptian formula once it is understood that $(a-b)(a^2+ab+b^2) = a^3 - b^3$. The consensus is that the Egyptians were aware of this identity..

The later Egyptian geometry seems to have been influenced by the Babylonians since on the tomb of Ptolomy XI who died in 51 BC the inscription contained the incorrect formula for the area of a quadrilateral of sides a, b, c and d

$$\frac{(a+c)(b+d)}{4}.$$

The Babylonians and the Egyptians also had an understanding of the geometry of circles.

2.1.4 Exercises.

1. Can you find any quadrilaterals have area in accordance with the Babylonian formula?

2. What is the area of a rectangular trapezoid with dimensions 2,4,3?

3. A problem on the Moscow Papyrus says: One leg of a right triangle is two and a half times the other and the area 20. What are its dimensions? Use the Egyptian method of completing to a rectangle to solve the problem (this is the way it was done on the papyrus).

4. Calculate the volume of a right pyramid (not truncated) of base 9 and height 15.

2.2 Thales and Pythagorus.

The geometry of the ancient civilizations is important but pales next to the developments in early Greece. Perhaps one reason why we are so aware of Greek mathematics is because of their rich literature and historical writing dating from the earliest eras of their civilization. The first Olympic games were held in 776 BC (a documented historic event). The works of Homer and Hesiod (still read) predate this event. During the sixth century BC there is a record of two great mathematicians Thales and Pythagorus. Their individual achievements are only documented by secondary sources which, perhaps, exaggerate the accomplishments of these two mathematicians.

The Greek world in 600 BC had spread from its original boundaries of the Aegean and Ionian seas to scattered settlements along the Black and Mediterranean Seas. Most of the mathematics that has been recorded comes from these outskirts. One possible reason for this is that they interacted with the older cultures of the Babylonians and the Egyptians. Thales of Milatus (624-548 BC) and Pythagorus of Samos (580-500 BC) were known to have travelled to the ancient centers of Babylonia and Egypt to study their mathematics.

2.2.1 Some theorems of Thales.

Eudemus of Rhodes (320 BC a student of Aristotle) wrote a history of mathematics that is now lost but a summary of this history (also lost) was incorporated by Proclus (410-485 AD) in his early pages of commentary on the first book of the Elements by Euclid. Proclus reports as follows:

... (Thales) first went to Egypt and thence introduced this study to Greece. He discovered many propositions himself and instructed his successors in the principles underlying many others, his methods of attack being in some cases more general in others more empirical.

Later quoting (the quote of) Eudemus he attributes that following five theorems (found in the Elements) to Thales.

1. *A circle is bisected by its diameter.*

2. *The base angles of an equilateral triangle are equal.*

3. *If two lines intersect the two opposite angles are equal.*

4. *If two triangles have all their angles equal then the corresponding sides are in proportion.*

5. *If two triangles have one side and the two adjacent angles equal then they are equal.*

We will consider these theorems in our discussion if Euclidean geometry. Thales was a practical man whose motto according to Proclus was ”know thyself”.

2.2.2 Pythagorus.

Pythagorus, on the other hand, was a mystic and a prophet. His motto (and that of the Pythagoreans) was “all is number”. As with Thales, only secondary sources still exist (Aristotle was known to have written a biography of Pythagorus). The Pythagoreans were a vegetarian sect since they believed (possibly influenced by a trip of Pythagorus to India) in the migration of souls. The term mathematics comes from Pythagorus and literally means “that which is to be learned”. Proclus, in his introduction to the books of Euclid says:

Pythagorus, who comes after him [Thales], transformed this science into a liberal form of education, examining its principles from the beginning and probing the theorems in an immaterial and intellectual manner[meaning abstract]. He described the theory of proportions and the construction of cosmic figures.

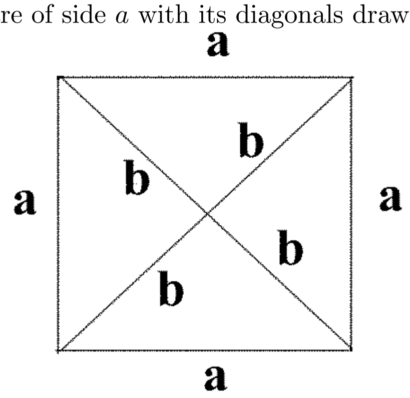
Johann Stevin Kepler (1571-1630) wrote:

Geometry has two great treasures: one is the Theorem of Pythagorus; the other, the division of a line into extreme and mean ratio. The first we may compare to the measure of gold the second we may name a precious jewel.

The meaning of this quotation will be clearer after reading the next section.

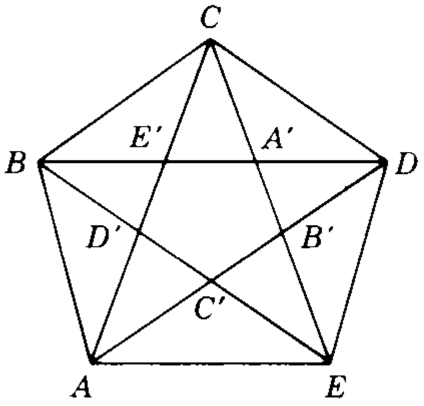
2.2.3 The golden ratio.

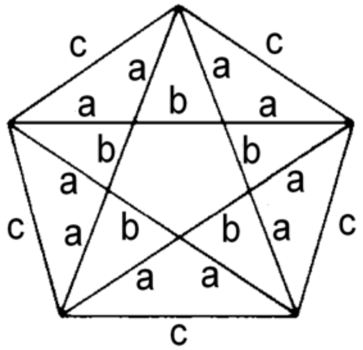
First we look at a square of side a with its diagonals drawn.



This picture is quite similar to one on the Babylonian tablet Yale 7289. We note that if we use the method the Babylonians each of the triangles with legs b,b and hypotenuse a has area $\frac{b^2}{2}$ since 4 of them make up the square of area a^2 we see that $a^2 = 4\left(\frac{b^2}{2}\right) = 2b^2$. Note that the obvious symmetry implies that all of the angles in the center are equal and so each must be a right angle. The drawing is therefore an elegant proof of the Pythagorean theorem in this case.

It is reasonable to ask what happens for a pentagon? Consider the two figures





If we rotated the pentagon so that a vertex would go to a vertex then the figure would look exactly the same. This says that all the segments labeled by an a are equal to the same value which we will call a. Similarly for the ones marked b and c. Now each of the triangles with base c and two sides a full diagonal $(a + b + a)$ rotate one on the other. For example, AEC and ABD. Each of the triangles with base b and sides a (for example, $E'A'C$) is similar to the of the triangles with base c and sides $a + b + a$. So

$$\frac{2a+b}{c} = \frac{a}{b}.$$

We note that the line BD is parallel to AE and that AC is parallel to ED. This implies that $AE^{'}$ has the same length as ED. So $c = a + b$. We therefore have

$$\frac{2a+b}{a+b} = \frac{a}{b}.$$

Cross multiplying gives $2ab + b^2 = a^2 + ab$. Hence

$$b^2 + ab = a^2$$

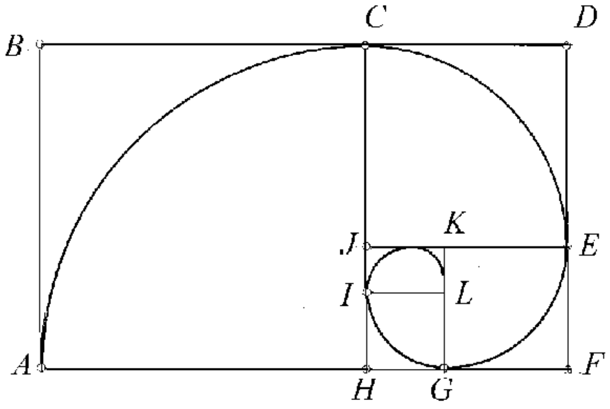
If we divide both sides of this equation by a^2 and set $x = \frac{b}{a}$ then the equation becomes

$$x^2 + x = 1.$$

Thus the ratio $\frac{b}{a}$ satisfies the above equation. This ratio was called the *golden ratio* by the Pythagoreans. It is this division that Kepler called a precious jewel. The ancients believed that a rectangle whose sides are in this ratio was the most pleasant to the eye. The Greeks designed the Parthenon so that its sides conformed to this ratio. The number $y = \frac{1}{x}$ is called the *golden section* and satisfies

$$y^2 = 1 + y.$$

A rectangle whose smaller side is in the proportion of the golden ratio to the larger is called a *golden rectangle.* It has the property that if we take a golden rectangle $ABDF$ as in the picture below and if we We mark the point C so that the length of BC equals the length of the shorter side AB. Then one has a subdivision into a square $ABCH$ and a rectangle $CDFH$. The rectangle is another golden rectangle:



To see this we observe that if b is the length of AB and if a is that of BD then CD has length $a - b$ and FD has length b. We assert that

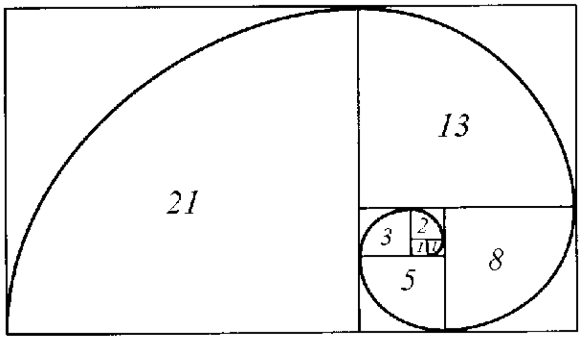
$$\frac{a-b}{b} = \frac{b}{a}.$$

To see this cross multiply we are trying to see if $a(a-b) = b^2$. That is if $a^2 = ab + b^2$. Which is just the assertion that $\frac{b}{a}$ is the golden ratio. The point here is we can now look at the new golden rectangle as a square and a golden rectangle. In fact, we can continue this forever. If in each of the squares we draw the part of the circle of radius equal to the length of a side starting at the far corner (relative to our labeling) we have a spiral. The arcs seem to fit smoothly. They do (but not as smoothly as the picture indicates) and we'll understand why after we discuss the infinitesimal calculus.

2.2.4 Relation with the Fibonacci sequence.

If we recall the problem of Fibonacci: A rabbit takes one month from birth to become sexually mature. Each month a mature pair gives birth to two (assume a male and a female) rabbits. If you have a pair of newborn rabbits how many pairs rabbits will you have in a year? We start with 1 pair after a month they have just gotten mature so there is still only one pair. They give birth at the end of the next month so there are then 2 pair. In one month the original pair will give birth again but the second pair will have just gotten sexually mature. So there are 3 pairs. 2 of the pairs sexually mature and 1 newborn. The next month the two pairs of mature rabbits give birth and the newborn matures there are now 5 pairs of rabbits 3 mature and 2 newborn. Next month there will be 5 mature and 3 newborn. The pattern (first apparently pointed out by Kepler) is if the number of rabbits at the beginning of month k is denoted F_k with N_k newborn and M_k mature then $M_{k+1} = F_k$ (every rabbit that existed at the beginning of month k is mature in one month) and $N_{k+1} = M_k$ (only the mature give birth in one month). Since $F_{k+1} = M_{k+1} + N_{k+1}$, we see that $F_{k+1} = F_k + M_k = F_k + F_{k-1}$. Viz.. $F_0 = N_0 = 1$ (this is where we start), $F_1 = M_1 = 1$, $M_2 = 1$,$N_2 = 1$ so $F_2 = 2$. Now $F_3 = F_2 + F_1 = 3$. Similarly, $F_4 = F_3 + F_2 = 3 + 2 = 5$. Continuing in this way we have the sequence

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144....$$



If we start with two squares of side one. Put a square of side two on top of them. Then a square of side 3 to the left. After that a square of side 5 below, etc. (as in the picture) .If we draw circles as in the case of the golden spiral we find that we have a spiral that is almost identical. This leads us to consider the ratios $\frac{F_k}{F_{k+1}}$ we have

1, .5, .666..., .6, .625, .6153846, .61904 76, .6176471, .61818..., .6179775, .618055...

for the first 12 ratios to at least an accuracy of 7 digits. We note that the golden ratio is .61803398875 to 11 decimal places. This seems to indicate that if we us the notation α for the golden ratio and if we write $Q_k = \frac{F_k}{F_{k+1}}$ then

1. $Q_{2k-1} < \alpha < Q_{2k}$ for $k = 1, 2, 3, ...$
2. $Q_{2k} - Q_{2k-1}$ becomes arbitrarily small with k.
3. $Q_{2k-1} < Q_{2k+1}, Q_{2k} > Q_{2k+2}$.

If these observations are true then the Fibonacci sequence gives an effective way of calculating the golden ratio to arbitrary precision. These observations are true (as we shall soon see) but a much more surprising relationship is true. We first observe that the quadratic formula implies that

$$\alpha = \frac{\sqrt{5}-1}{2}.$$

We also note that if we use the notation τ for the golden section then

$$\tau = \frac{\sqrt{5}+1}{2}.$$

A theorem of J.P.M.Binet proved in 1843 says

$$F_k = \frac{\tau^{k+1} - (-\alpha)^{k+1}}{\sqrt{5}}, k = 0, 1, 2, ...$$

Let us check this for some small k. If $k = 0$ then the numerator is $\frac{\sqrt{5}+1}{2} + \frac{\sqrt{5}-1}{2} = \sqrt{5}$. So the formula is correct. If $k = 1$ then the numerator is $(\frac{\sqrt{5}+1}{2})^2 - (\frac{\sqrt{5}-1}{2})^2 = \sqrt{5}$. To prove the formula for all k we will use mathematical induction. The assertion S_n is that the formula is true for all k between 0 and n. We know that S_0 and S_1 are true. Let us assume that S_n is true. We must show that S_{n+1} is true. To do this we need only show that the formula is correct for F_{n+1} and we may assume that $n + 1 \geq 2$. Thus

$$F_{n+1} = F_n + F_{n-1}.$$

Our assumption implies that

$$F_n = \frac{\tau^{n+1} - (-\alpha)^{n+1}}{\sqrt{5}}, F_{n-1} = \frac{\tau^n - (-\alpha)^n}{\sqrt{5}}.$$

If we add these two terms together we find that

$$F_n + F_{n-1} = \frac{\tau^n(\tau + 1)}{\sqrt{5}} - \frac{(-\alpha)^n(1 - \alpha)}{\sqrt{5}}.$$

We now observe that $\alpha^2 = 1 - \alpha$ and $\tau^2 = \tau + 1$, So

$$F_n + F_{n-1} = \frac{\tau^n\tau^2}{\sqrt{5}} - \frac{(-\alpha)^n\alpha^2}{\sqrt{5}} = \frac{\tau^{n+2} - (-\alpha)^{n+2}}{\sqrt{5}}.$$

This is the desired formula for F_{n+1}. Notice that we have given no indication as to why we thought that such a theorem might be true. The method of mathematical induction can only be used to prove assertions that we have guessed in advance or, perhaps, we can derive in using deeper insight. We will describe an alternate more direct approach when we study matrix theory.

The formula of Binet easily implies the 3 assertions above. We have (using $\alpha = \frac{1}{\tau}$) that if k is even than

$$Q_k = \frac{\alpha + \alpha^{2k+3}}{1 - \alpha^{2k+4}} > \alpha$$

if k is odd then

$$Q_k = \frac{\alpha - \alpha^{2k+3}}{1 + \alpha^{2k+4}} < \alpha.$$

Also

$$Q_{2k} - Q_{2k-1} = \frac{\alpha^{4k}(\alpha + 2\alpha^3 + \alpha^5)}{(1 + \alpha^{4k+2})(1 - \alpha^{4k+4})}.$$

This implies 2. We leave 3. to the reader.

We note that Binet's formula can be used to prove many results about the Fibonacci sequence. One very nice formula is

$$F_{n+1}F_{n-1} - F_n^2 = (-1)^{n+1}.$$

To check this we do the substitution of Binet's formula in the left hand side of the equation:

$$F_{n+1}F_{n-1} - F_n^2 = ((\tau^{n+2} - (-\alpha)^{n+2})(\tau^n - (-\alpha)^n) - (\tau^{n+1} - (-\alpha)^{n+1})^2)/5.$$

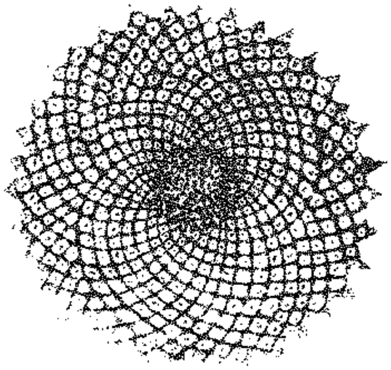
If we multiply out the terms in the braces the left hand side of this equation is equal to

$$\frac{-\tau^{n+2}(-\alpha)^n - (-\alpha)^{n+2}\tau^n + 2\tau^{n+1}(-\alpha)^{n+1}}{5}.$$

We now use the facts that $\alpha\tau = 1$ and $\tau^2 + \alpha^2 = 3$. So the above display is indeed $(-1)^{n+1}$.

2.2.5 Phyllotaxies.

In this subsection we will discuss an apparent relationship between the Fibonacci numbers and the spiraling that occurs in plants. It has been observed that the number of petals of a specific type of flower is usually a Fibonacci number. Lilies have 3, buttercups 5, marigolds 13, asters 21 most daisies 34,55, or 89. The head of a flower (like a sunflower or a daisy) can be seen to have two families of interlaced spirals, one winding clockwise and the other counterclockwise. The pair of numbers is (see the figure below) 34 and 55 or 55 and 89 for the sunflower. Another such phenomenon is the spiralling of pine cones. Among the pine cones found in a cursory look in Del Mar, California one can find 5,8 and 8,13 pine cones.



There are many attempts at an explanation as to why the Fibonacci numbers appear in so many ways in nature. The most convincing are related to the assertion that the "golden ratio is the most irrational number". We will give one explanation of this statement in terms of continued fractions.

First we will explain the idea of a continued fraction. If we have a number $0 < a < 1$ then an expression for a as a continued fraction is

$$a = \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \ldots}}}$$

With $a_1, a_2, \ldots$ positive integers. This means that we should consider

$$\frac{1}{a_1}, \frac{1}{a_1 + \frac{1}{a_2}}, \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3}}}, \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \frac{1}{a_4}}}}, \ldots$$

as better and better approximations to a. These rational numbers are called the *convergents* (here we have the first,second,third and fourth convergent). Here is the method for finding $a_1, a_2, \ldots.$ Define $r_1 = \frac{1}{a}$, and a_1 to be the largest integer less than or equal to r_1. In general, assuming r_n and a_n have been defined and $r_n > a_n$ then define

$$r_{n+1} = \frac{1}{r_n - a_n}$$

and a_{n+1} to be the largest integer less than or equal to r_{n+1}. If $r_n = a_n$. Then the nth convergent is equal to a and we stop. If a is irrational this procedure will never stop. If $a \geq 1$ then we set a_0 equal to the largest integer less than or equal to a and we write $a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \ldots}}}$ for the continued fraction and the convergents are

$$a_0 + \frac{1}{a_1}, a_0 + \frac{1}{a_1 + \frac{1}{a_2}}, a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3}}}, a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \frac{1}{a_4}}}}, ...$$

Then one can show that if $\frac{p_n}{q_n}$ is the nth convergent then $\frac{p_n}{q_n}$ is in lowest terms and is closer to a then any fraction in lowest terms with denominator at most q_n. If a is the Golden Mean then $r_1 = \frac{2}{\sqrt{5}-1} = \frac{\sqrt{5}+1}{2}$. Thus $1 < r_1 < 2$ so $a_1 = 1$.

$$r_2 = \frac{1}{\frac{\sqrt{5}+1}{2} - 1} = \frac{2}{\sqrt{5}-1}$$

thus $r_2 = r_1$ so $a_2 = a_1 = 1$. We note that this goes on forever, $r_{n+1} = r_n$ for $n = 1, 2, ...$ and thus $a_{n+1} = a_n = ... = a_1 = 1$. Thus the partial fraction expansion of the Golden Mean is

$$\frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + ...}}}}.$$

The convergents are

$$1, \frac{1}{2}, \frac{2}{3}, \frac{3}{5}, \frac{5}{8}, \frac{8}{13}, ...$$

Which we recognize as $\frac{F_{n-1}}{F_n}$ for $n = 1, 2, 3, ...$

Recent explanations of phyllotaxies involve this irrationality and these convergents. The suggested theory is that if “blobs of expanding matter” that radiate from a central source are such that as they radiate out they repel then they should be propelled initially at a slope that is badly approximated rationally thus allowing the most space for the radiated initial blobs which are all assumed to be the same size. The fact that the golden ratio is so badly approximated makes it a likely candidate for this angle of radiation. The corresponding ratio of the counts of spirals then give a rational approximation to this number. One of the first observations of this phenomenon is in the work of 1837 Auguste and Louis Bravais who observed this angle in the ratio of the left and right spiraling of leaves on many trees. In 1872 P.G. Tait extended the work of the Barvais brothers to the spiralling we have been discussing. A controlled experiment was performed by Stépane Douady and Yves Couder in 1993(La Reserche, 24 (1993), 26-25) which confirms these ideas. In their experiment they had a medium of liquid silicon on a disk and from the center of the disk they “shot” blobs of magnetized liguid. On the edge of the disk they had a strong magnetic source which would cause the blobs to radiate. They found that the count of the spiraling depended on the rate of radiation. The most likely count was a pair of consecutive Fibanocci numbers. However, by changing the rate they found other sequences such as $1, 3, 4, 7, 11, ...$ This sequence satisfies the same recursion as the Fibonacci sequence. An amusing discussion of this work can be found in the Mathematical Recreations column of Ian Stewart in the January 1996 Scientific American.

2.2.6 Exercises.

1. Join the vertices of a regular six sided figure. Can you see any interesting ratios, etc.?

2. Explain why the golden spiral looks smooth.

3. Show that $Q_{k+1} = \frac{1}{Q_k+1}$. Use this to show assertion 3. above directly. Suppose that we have a sequence a_k of positive rational numbers satisfying $a_{k+1} = \frac{1}{1+a_k}$. Show that if $\lim_{k \longrightarrow \infty} a_k$ exist then the limit is α.

4. Use mathematical induction to show that $\tau^{n+1} = F_n\tau + F_{n-1}$, $n = 1, 2, ...$ What is the analogous formula for α?

5. We define a sequence $E_0 = 1, E_1 = 1$ and $E_{n+1} = E_0 + E_1 + ... + E_n$. What can you say about this sequence?

6. Consider the sequence defined by the following rules, $A_0 = 3, A_1 = 0, A_2 = 2, A_{n+1} = A_{n-1} + A_{n-2}$. This sequence is called the Perrin sequence. In 1991 Steven Arno proved that if n is prime then n divides A_n. ($A_3 = 3$, $A_4 = 2$, $A_5 = 5$, $A_6 = 5$, $A_7 = 7$, $A_8 = 10$, $A_9 = 12$, $A_{10} = 17$, $A_{11} = 22, ...$). It has been shown that calculating the remainder of the division of A_n by n is easy (in the sense of section 7 of Chapter 1). Devise a primality test based on this result.

7. Use the formula $F_{n+1}F_{n-1} - F_n^2 = (-1)^{n+1}$ to show that consecutive Fibonacci numbers are relatively prime.

8. Find as many examples (or counter examples) to the phenomenon described in the above section (phyllotaxies).

9. Show that the nth convergent of the Golden Mean is $\frac{F_{n-1}}{F_n}$.

10. Let $a = \pi$. Show that the 0th convergent is 3 and the first is $\frac{22}{7}$ (you can use 3.1416 as an approximation for π).

2.3 The Geometry of Euclid.

When we think of the work of Euclid we think about his Thirteen Books of the Elements and plane geometry. We have already seen that this is a misconception. Books VII,IX and XI are concerned with number theory. Solid geometry also appears in several places Books X,XI,XII and XIII. He also wrote books on other topics. Some of his work still exists including his Optics and a book called Phenomena which is a treatise on spherical geometry as it applies to astronomy. He also wrote The Elements of Music which is unfortunately lost. However, his book Sectio Canonis on the Pythagorean theory of music still exists. Without a doubt, his reputation rests on his masterpiece: The Elements. Since the geometry in the elements is much better known than the number theory, we will make an even less complete study of it than we did of the number theory. As in the case of the number theory Book I begins with definitions 23 in this case. There are then 5 Postulates and 5 Common Notions.

2.3.1 The definitions.

1. A *point* is that which has no part.

Like his first few definitions in Book VII this definition must be taken with a grain of salt. He seems to mean that points are the smallest objects that we will consider.

2. A *line* is breadthless length.

As we shall see a line is not necessarily a straight line. In fact, we will see an attempt in Definition 3 to define a straight line. In modern terminology Euclid's line would be a curve. (Definition 15 defines a circle as a part of a line.)

3. The extremities of a line are points.

4. A *straight line* is a line which lies evenly with the points on itself.

This is Euclid's expression for a line as we know it. It seems clear that he is asking us to picture a straight line and is just saying that our picture is correct. In a nutshell, a straight line is a line that has some sort of uniformity that should imply straightness.

5 defines a surface, 6 says that the extremities of a surface are lines and 7 defines a plane surface. These definitions are completely analogous to what he does for lines and straight lines.

8. A *plane angle* is the inclination to one another of two lines that meet each other and do not lie on a straight line.

Here he is giving us the notion of an angle between (what we would call two curves). He doesn't seem to think that a definition of inclination is necessary. Furthermore he must be thinking of lines that have exactly one point in common (where they meet) but both do not lie on the same straight line. This is a bit confusing since the lines are not necessarily straight. We can conceive of curves that are partially in a straight line and partially off of it. With the use of the methods of Calculus one can give a notion of angle between two curves. But these curves must be well approximated by straight lines near the point where they meet.

9. And when the lines containing the angle are straight, the angle is called *rectilineal.*

This is defining what we usually mean by an angle (that is between two straight lines). Next he defines a right angle.

10. When a straight line set up on a straight line makes the adjacent angles equal to one another, each of the equal angles is *right* and the straight line standing on the other is called a *perpendicular* to that on which it stands.

Here we are asked to know what it means for two angles to be equal. Euclid seems to have no need to define such a concept. It seems clear that he feels that he must introduce some terminology but that all he is doing is describing objects with which we are already familiar. The next definitions define obtuse angle to be one greater than a right angle and acute angle to be one less than a right angle. Euclid does not seem to feel that he has any need to explain the meaning of the terms less than or greater than in the context of angles. Definitions 13 and 14 are concern boundary and figure. A boundary is defined to be an extremity but an extremity in this context is not defined. Although there is an indication in Definition 3. What he seems to mean is that the boundary is swept out by extreme points of lines. A figure is that which is contained in a boundary.

15. A *circle* is a plane figure contained by one line such that all straight lines falling upon it from one point among those lying within the figure are equal to one another.

16. And the point is the *center* of the circle.

So a circle is contained by one line. So a line is really what we think of as a curve. There is a point so that if we take a straight line with one extremity at this point and the other on the circle getting a straight line L then do the same for another point on the circle getting a straight line M and if we lie the two lines one on top of the other they are the same. Definitions 17 and 18 define diameter and semicircle. We note that one of the parts of the definition of diameter is the first Theorem that Eudemus attributed to Thales. We should also note that Euclid felt no need to prove this part of the definition. 19,20,21,22 define various types of figures using the terminology with which we are all familiar. Definition 23 involves a concept that is needed in the statement of the fifth Postulate.

23. *Parallel* straight lines are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

The point of the definitions seems to be to attach names to concepts that we already know. Euclid's definitions are not definitions as they are understood in modern mathematics.

2.3.2 The Postulates.

Here Euclid describes assumptions that he feels must be made as the basis of geometry. These are of two types. The first 3 describe constructions that are possible.

1. To draw a straight line from any point to any point.

In other words if we have two points there is always a straight line that joins them.

2. To produce a finite straight line continuously in a straight line.

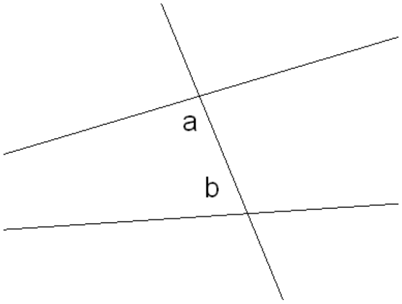
This can mean several things. He seems to want it to mean that we can choose any point on a straight line and have that point be one of the endpoints of a straight line of fixed length.

3. To describe a circle with any center and distance.

We can draw a circle with any center and any radius (in any plane).

The next two are assertions about angles.
4. That all right angles are equal.

5. That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on the side on which are the angles less than the two right angles. (In the picture below the angles in question are a and b.



This is the famous parallel postulate. It seems obviously true and is confirmed by every picture we draw. We will see that it was the subject of intense speculation into the nineteenth century. The brunt of the study was to see if it could be deduced from the other 4 using the definitions and the common notions that we will now describe. This work on the parallel postulate will be studied in greater detail after we have developed a more sophisticated groundwork for our analysis.

2.3.3 The Common Notions.

These are the basic axioms for equality and inequality.

1. Things that are equal to the same thing are equal to each other.

2. If equals be added to equals the wholes are equal.

This common notion is a geometric assertion. It applies to areas, geometric figures and numbers (as in the definitions before Book VII). The next common notion should be interpreted in this way also.

3. If equals be subtracted from equals the remainders are equal.

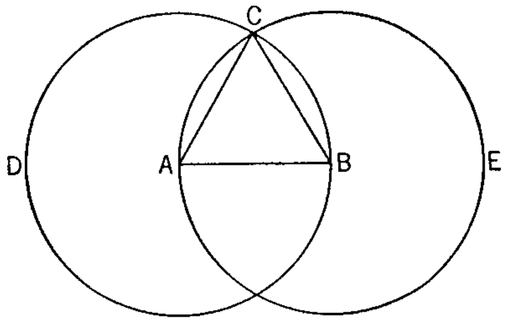
4. Things which coincide with one another are equal.

This is the basic method of showing that things are equal in the Elements. The proofs devise a method of laying one object onto another object in such a way that they coincide. That is they fit together perfectly. This can be seen graphically in Proposition 4 of Book I which shows that if two triangles have to pairs of equal sides and the included angles are equal then if you lay the angle made by the corresponding sides of one triangle onto that for the other the two triangles coincide.

5. The whole is greater than the part.

2.3.4 Some Propositions.

Euclid is now in business. All terms he will need in Book I are defined (we should assume to his satisfaction also other books such as book II will define more terms). The rest of Book I involves basic plane geometry. We give the flavor of the proofs by looking at two examples, in detail. Proposition 1 and Proposition 47 (The Pythagorean Theorem) in Book I. We will first look at Proposition 1.



On a given finite straight line construct an equilateral triangle.

This means that we are asked to show that if we are given a finite straight line (an interval) we can construct an equilateral triangle with one side equal to the given one. We will now give the proof as given in Euclid

The argument is as follows. We have the line AB. We use Postulate 3 twice to make the two circles shown the first with center A the second with center B and both with distance AB. Let C be the intersection of the two circles. Then $AC = AB$ by the definition of circle (Definition 15) and $BC = AB$ for the same reason (that AC and BC exist is Postulate 1). Thus $AC = BC$ by 1. in the Common Notions. The triangle thus has all of its sides equal.

This is fine except for one assertion that Euclid does not deem necessary to be proved: That the circles intersect. This is more serious than the lack of the need to prove what we called the Euclidean property in section 1.4. A proof of the existence of this intersection involves more sophisticated mathematics. At a minimum it involves real definitions of some of the terms. The crux of the matter has to do with the fact that a circle has an inside and an outside and that a line (or a circle) that contains a point in the inside of the circle and a point on the outside must have a point on the circle itself.

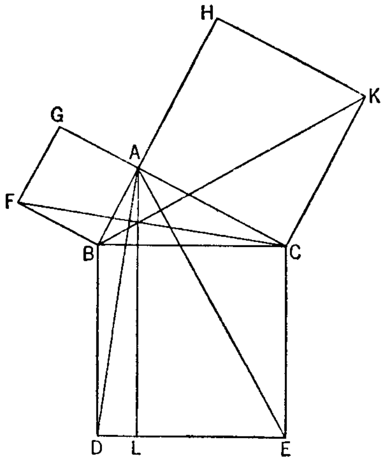
Proposition 5 is the assertion that the base angles of an isosceles (legs equal) triangle are equal. This is a strengthening of Thale's second Theorem as we quoted from Eudomus.

Proposition 15 is the third Theorem of Thales that we quoted.

Proposition 26 is the fifth Theorem of Thales in our list.

The other Proposition that we will analyze in detail is number 47 in Book I. We call it the Pythagorean theorem. The proof below seems to be original to the Elements (in other words most of the other proofs are transcriptions of other people's arguments).

In right-angled triangles the square of the side subtending the right angle is equal to the squares on the sides containing the right angle.



The basic idea is to show that the triangles ABD and FBC are equal as are the triangles AEC and BCK. To see how this proves the theorem we note that since the triangle ABD has base BD and height DL as does the rectangle with sides BD and DL (Euclid simply calls it the parallelogram BL). We conclude (as did the Egyptians, Babylonians and Proposition 41, Book I) that the rectangle BL is twice the triangle ABD. The same argument shows that the square $ABFG$ is twice the triangle BCF. Hence since doubles of equals are equal to each other (this is a statement in braces without any further reference) this implies that the square $ABFG$ is equal to the rectangle BL. Similarly, the square $ACKH$ is equal to the rectangle CL. Since, BL and CL make up the square $BCED$ the Proposition follows. (Euclid says: Therefore etc. Q.E.D.).

We are left with the assertion about the triangles. We will consider the first pair notice that $AB = BF$, $BD = BC$ thus in light of Proposition 4 Book I

(Thales fifth theorem in our list above) we need only show that the angles ABD and FBC are equal. To see this we observe that the angle DBA is the sum of DBC and ABC. The angle FBC is the sum of ABF and ABC. Since all right angles are equal $ABF = DBC$. So the assertion about the angles follows from Common Notions 2.

2.3.5 Exercises.

1. Prove the converse of the Pythagorean theorem. That is, if the square of one side of a triangle is equal to the sum of the squares of the other two sides then the angle opposite this side is a right angle. (This is Proposition 48 in Book I. Explain the proof in the Elements and give a proof using, say, trigonometry).

2. In Proposition 11 of Book II of the elements show that Euclid is showing that one can construct the Golden ratio.

2.4 Archimedes.

Archimedes lived during the period 287-212 BC. He was a citizen of Syracuse. In his youth he is thought to have traveled to Egypt and while there he invented the water screw as a way of lifting large amounts of water. He developed a theory of levers and made the famous boast : "Give me a place to stand on and I can move the Earth." He is said to have backed this up by raising a ship out of the water using one arm. He was also a military engineer who invented many weapons during the defense of Syracuse against the Romans. He is said to have used giant lenses to focus the sunlight to burn down the Roman fleet. The history of his practical inventions is largely second hand since he wrote commentary on only one of these (*On sphere making* which is lost).

Most of Archimedes' writings on mathematics have been preserved. He wrote his work in the form of letters to his friends: Conon of Samos and Eratosthenes. After Conon died he sent his letters to Conon's student Dositheus of Pelusium. When the Romans eventually invaded Syracuse in 212 BC their general, Marcellus, ordered that Archimedes and his household be spared in the ensuing massacre. However, when a soldier went to escort Archimedes to an audience with Marcellus, Archimedes was concentrating on a geometric problem. He told the soldier that he would come once he solved the problem. The soldier was furious and killed Archimedes: perhaps the greatest mathematician who ever lived.

At this point we will discuss two of Archimedes works: *The Sand-Reckoner* and *Measurement of the Circle*. The first is in the nature of a study of very large numbers. The second is the genesis of the elegant approximation $\frac{22}{7}$ of π. In later chapters we will be looking at Archimedes work on what we call calculus (although the work alluded to above on the calculation of π involves ideas that are usually associated with calculus.

2.4.1 The Sand-Reckoner.

This paper begins with an introduction written in the form of a letter to King Gelon of Syracuse and ends with a conclusion also addressed to Gelon. Let us quote from the initial material in the paper (as translated by Heath).

“There are some, King Gelon, who think that the number of the sand is infinite in multitude; and I mean by sand not only that which exists about Syracuse and the rest of Sicily but also that which is found in every region whether inhabited or uninhabited. Again there are some who, without regarding it as infinite, yet think that no number has been named that is great enough to exceed its multitude...But I will try to show you by geometrical proofs, which you will be able to follow, that, of the numbers named by me and given in the work I sent to Zeuxippus, some exceed ... that of the mass equal to the magnitude of the universe.” [Here the point that will be critical is the phrase ”no number has been named that is great enough...”]

He then discusses the various possibilities for the size of the universe with the idea that whatever sizes are believed he will always take one bigger. Included in the models that he considers is that of Aristarchus of Samos of which Archimedes says: “His hypothesis are that the fixed stars and the Sun remain unmoved, that the Earth revolves about the Sun in the circumference of a circle, the Sun lying in the middle of the orbit, and that the sphere of the fixed stars, situated above the same center as the sun, is so great that the circle in which he supposes the Earth to revolve bears such a proportion to the distance of the fixed stars as the center of the sphere bears to its surface.” Archimedes goes on to discount this theory for technical reasons. However, his point is not to establish a theory of the universe but just to get an upper bound on its size. Now comes the point of the whole exercise. There was no known notation or theory of big numbers. Recall that the Egyptians really didn’t get past 10 million. The Romans would be constantly inventing new symbols and would eventually run out of letters in the alphabet. The biggest number that the Greeks used was a myriad which is 10,000. Archimedes considers what happens if we multiply two myriads. One then has a myriad myriads. Then he proposes to take a myriad myriads and treat it as a basic unit (a number of the first order) then he can multiply it by a myriad myriads. One can continue this way a myriad number of times and get a number that Archimedes called P (probably π but we reserve this symbol for something else) a number of the second order. In modern notation $P = (100000000)^{100000000}$. He then observes that he can continue this process by taking P to be a number of the first order and consecutively multiply P by itself P times getting $P^P = (100000000^{100000000})^{100000000^{100000000}}$. This new number can now be treated as a number of the first order and the process repeated once again. He then gives a reasonable argument that one of his new found immense numbers is big enough. He in fact argues that the number of particles in the universe is less than 10^{63} (much smaller that P). The modern estimates are somewhat nearer to 10^{100}.

In modern mathematics we the ideas of this paper lead to the *Archimedian Property* that is given any number (we mean here a rational or real number) there is an integer that is strictly bigger.

2.4.2 Exercises.

1. Prove the following is a theorem in The Sand-Reckoner by induction (this is the way Archimedes proved it):

If there be any number of terms in continued proportion say $A_1, A_2, ..., A_n, ...$*and if the first is* 1 *the second is* 10 [so the third is 100] *and if the* mth *term is multiplied by the* nth *term the distance* [i.e. the number of terms between them] *from this term to* A_n *is the same as the distance from* 1 *to* A_m.

2. In Archimedes paper he takes as the diameter of the Sun 30 times the diameter of the moon. Do you agree with this? (He quotes Euxedus, his own father Pheidias and Aristarchus for estimates of 9, 12 and 20 times. Thus he was estimating higher than anyone else at the time.)

2.4.3 Archimedes' calculation of π.

We will next look at Archimedes' study of the number π. He is the first to prove that $\frac{22}{7}$ is a remarkably good approximation to the ratio of the circumference of a circle to its diameter. In his paper *Measurement of a Circle* he in shows that

$$3\frac{10}{71} < \pi < 3\frac{1}{7}.$$

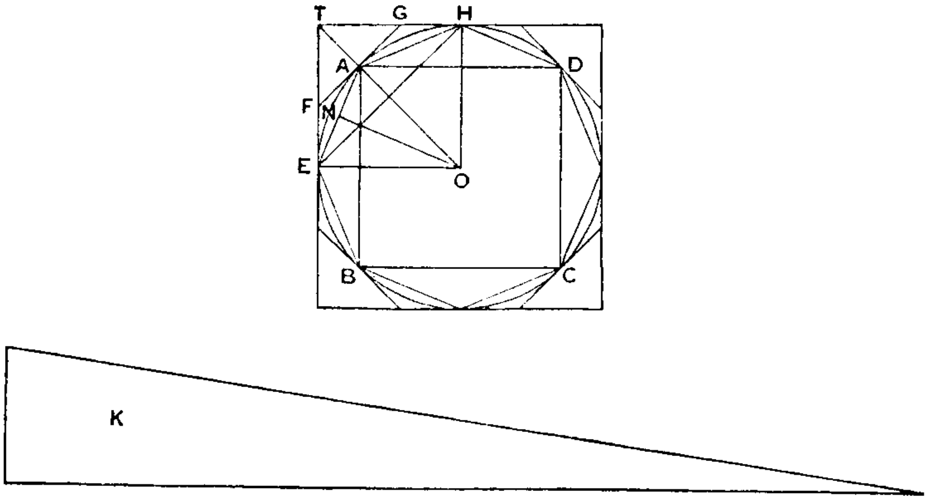
His method (as we shall see) could yield π to arbitrary precision. The important point to note is that he has lower and upper bounds of (in decimal notation) 3.1408 and 3.1429 thus π is 3.14.. to an accuracy of at least 0.002. After we study this remarkable result we will look at various ramifications of Archimedes work that span about 2200 years. Before we begin his we will discuss the understanding of π before Archimedes did his work. The Babylonians routinely used the value 3 for the ratio of the circumference to the diameter of a circle however in some tablets the other values closer to and perhaps including $\frac{22}{7}$ were indicated . In the Rhind Papyrus the value was taken to be $3\frac{1}{6}$ and sometimes $\left(\frac{16}{9}\right)^2 = 3.16...$ At least once in the Bible (Revised Standard Version 1952 the King James Bible is a bit more poetic but has the same meaning) in 1 Kings 7-23 it says:

"Then he made the molten sea; it was round, ten cubits from brim to brim, and five cubits high, and a line of thirty cubits measured its circumference."

The ten cubits from brim to brim is the diameter and the circumference is thirty so the ratio is 3. One can argue that the Bible wouldn't bother with fractions. But even so 31 or 32 would be much closer to the correct value. It is interesting that the first convergent of the partial fraction expansion of π is $\frac{22}{7}$

(see section 2.2.5). This means that there is no better rational approximation with denominator less than of equal to 7.

We will now give a discussion of Archimedes method. The first proposition involves the following diagram.



The area of any circle is equal to a right-angled triangle in which one of the sides about the right angle is equal to the radius, and the other to the circumference of the circle.

This proposition says the if a circle has radius r and circumference c then the area is $\frac{rc}{2}$. We know this in a different way. We know that $c = 2\pi r$ so the proposition says that the area is πr^2. However, this result allows us to calculate the area without knowing π. The argument is truly ingenious. Let K denote the area of the triangle and let a be the area inside of the circle. Archimedes observes that there are three possibilities $K < a, K > a, K = a$. The point is to show that the first two possibilities cannot occur. He first assumes $a > K$ and shows that this leads to a contradiction. He draws the inscribed square $ABCD$. He then bisects the arcs AB, BC, CD and DA and draw the lines from the center of the circle through the bisectors. If necessary bisect again and continue until the area of the inscribed figure is greater than K. To see that (under the hypothesis $a > K$) this is possible since all we need do is take the subdivision so fine that the sum of the maximal distances from the sides of the figure to the circle is less than $a - K$. That this can be done is obvious from the picture and although the Archimedesmethod of determination of the subdivision involves an assertion equivalent with the desired one that is unproved. He now observes that it is easily seen that the area of each of the polygonal figures is less than K.. In fact the area is the sum of the triangles whose vertices are consecutive vertices of the figure and the center. The height of each of these triangles is less than r and the sum of all the sides is less than c. Thus the area is less than K. So the case $a > K$ is impossible. To show that the case $a < K$ is impossible he argues as above using circumscribed polygons (see the picture).

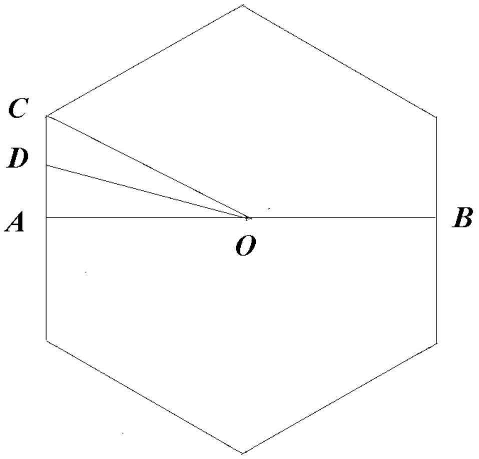
As we have pointed out there are still a few points in this argument that have not been proved (these are easily checked using trigonometry). However, if m is the area of *any* of the inscribed polygons as in the argument and if M is the area of any of the circumscribed polygons then we have

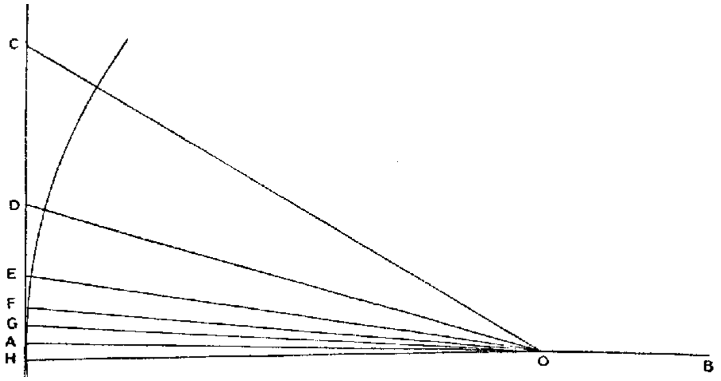
$$M > K > m$$

and

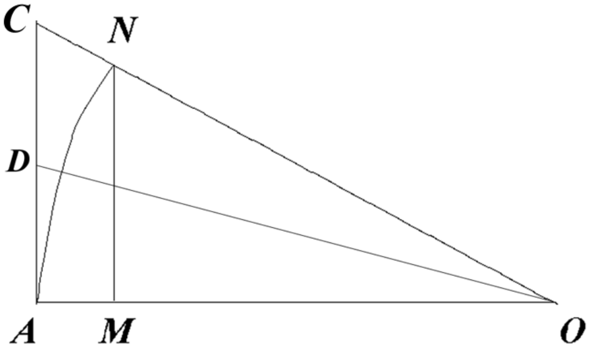
$$M > a > m.$$

To prove the result we must observe that for each (small) E there is a subdivision such that $M - m < E$. This is basically what Archimedes is asserting. Notice how close to modern calculus this is.



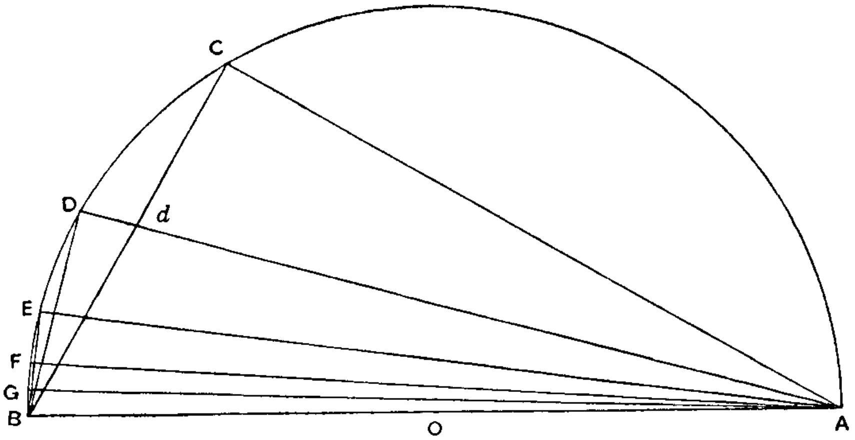


Archimedes next proves the upper bound for π using the above figure (the part OAC is to be thought of as part of the hexagon above it). In this figure the line BA is part of a diagonal of the circle. The line AC is tangent to the circle at A. He starts by taking the angle AOC equal to $\frac{1}{3}$ of a right angle. He then observes that $\frac{OA}{AC} > \frac{265}{153}$. Fortunately, another Greek mathematician named Eutocius inserted an explanation of this inequality. the actual ratio is $\sqrt{3}$. So we must just check that $\left(\frac{265}{153}\right)^2 < 3$. One checks that the square is 2.9999 to 4 decimal places. To see where the $\sqrt{3}$ comes from consider the following picture (the curve AN should be an arc of the circle of radius OA)



The angles OAC and OMN are right angles. The angle AOC is $\frac{1}{3}$ of a right angle so the angles ONM and OCA are each $\frac{2}{3}$ of a right angle. This implies that $\frac{NM}{ON} = \frac{1}{2}$ ($ON = OA$ since both are radii of the same circle). Now $ON^2 = NM^2 + OM^2$ so $OM^2 = \frac{3}{4}ON^2 = \frac{3}{4}OA^2$. Using the fact that OAC and OMN are similar triangles we see that $\frac{OM}{MN} = \frac{OA}{AC}$. Since $\frac{OM}{MN} = \frac{\frac{\sqrt{3}}{2}}{\frac{1}{2}} = \sqrt{3}$. The assertion follows. Similarly, $\frac{OC}{CA} = 2 = \frac{306}{153}$. Next we bisect the angle AOC which yields the line OD in the picture above. Now Proposition 3 in Book VI of the Elements implies that $\frac{CO}{OA} = \frac{CD}{DA}$ (see Exercise 1 in 2.4.4). Now we have $\frac{CO+OA}{OA} = \frac{CO}{OA} + 1 = \frac{CD}{DA} + 1 = \frac{CD+DA}{DA} = \frac{CA}{DA}$. Now multiplying the two ends of this string of equations by $\frac{OA}{CA}$ we have $\frac{CO+OA}{CA} = \frac{OA}{OD}$. We now use the inequalities $\frac{OA}{AC} > \frac{265}{153}$ and $\frac{OC}{CA} = 2 = \frac{306}{153}$. Thus $\frac{OA}{OD} = \frac{CO+OA}{CA} > \frac{265}{153} + \frac{306}{153} = \frac{571}{153}$. He now applies the Pythagorean theorem $OD^2 = AD^2 + AO^2$. So $\frac{OD^2}{AD^2} = \frac{AD^2+AO^2}{AD^2} > \frac{571^2+153^2}{153^2} = \frac{349\,450}{23\,409}$ ($153^2 = 23\,409$). Now Archimedes apparently guesses another very good lower bound for a square root ($(591+1/8)^2 = 3494\,28.7\,66...$) yielding $\frac{OD}{OA} > \frac{591\frac{1}{8}}{153}$. The point here is that he now has a good lower bound for the ratio $\frac{OD}{DA}$ instead of $\frac{OC}{CA}$. He can now bisect the angle AOD getting the point E in the main diagram above and argue as before getting a lower bound on $\frac{OE}{EA} > \frac{1172\frac{1}{2}}{153}$. He then bisects again and yet again. At his time he has an angle the size of $\frac{1}{48}$ of a right angle and at this fourth bisection he only does half the argument and gets an inequality $\frac{OA}{OG} > \frac{4673\frac{1}{8}}{153}$. Now the diameter of the circle is $2OA$ and $HG = 2AG$. Thus the ratio of the diameter of the circle to of the circumference of this 96 sided circumscribed regular figure is at least $\frac{4673\frac{1}{2}}{153\cdot 96}$. The reciprocally of this then gives an upper bound for π of $\frac{14688}{4673\frac{1}{2}} = 3 + \frac{667\frac{1}{2}}{4673\frac{1}{2}} < 3 + \frac{667\frac{1}{2}}{4672\frac{1}{2}} = 3 + \frac{1}{7}$.

The next task is to derive a lower bound for π. Archimedes does this by starting with a regular inscribed hexagon and bisecting 4 times just as he did for the upper bound.



He starts with the above picture with the angle BAC equal to one third of a right angle. As before, $\frac{AC}{CB} = \sqrt{3}$. This time Archimedes needs an upper bound for $\sqrt{3}$. He chose $\frac{1351}{780}$ since $\left(\frac{1351}{780}\right)^2 = 3.0000016...$ As before he bisects the angle BAC getting the straight line AD. He observes that AD intersects BD at the point d. We note that the angles at C and D are right angles. Thus since dAC and BAD are the two halves of the angle just bisected we see that the triangles ADB, ACd and dBD are similar. Thus $\frac{AD}{DB} = \frac{BD}{Dd} = \frac{AC}{Cd}$. Now we observe (see Exercise 1 below) that $\frac{CA}{AB} = \frac{Cd}{dB}$. Thus $\frac{AC}{Cd} = \frac{AB}{dB}$ (this implies that $(AC)(dB) = (AB)(Cd)$ which we will use in a moment). So $\frac{AD}{DB} = \frac{AB}{Bd}$. We also note that $(AB)(Bd) + (AB)(Cd) = (AB)(Bd) + (AC)(Bd)$ (see the parenthetic remark). Thus $\frac{AB}{Bd} = \frac{AB+AC}{Bd+Cd}$ (cross multiply). The denominator of the right hand side of this equation is $Bd + Cd = BC$. We therefor have $\frac{AB}{Bd} = \frac{AB+AC}{Bd+Cd} = \frac{AB+AC}{BC}$. The outgrowth of all of this is $\frac{AD}{DB} = \frac{AB+AC}{BC}$. We now note that $\frac{AC}{BC} < \frac{1351}{780}$ and $\frac{BA}{BC} = 2 = \frac{1560}{780}$. So $\frac{AD}{DB} < \frac{2911}{780}$. We now use the Pythagorean theorem for the right triangle BDA. Finding that $AB^2 = BD^2 + AD^2$. So $\frac{AB^2}{BD^2} = \frac{BD^2+AD^2}{BD^2} < \left(\frac{2911}{780}\right)^2 + 1 = \frac{2911^2+780^2}{780^2}$. As before Archimedes must approximate a square root and he takes an upper bound $\frac{AB}{BD} < \frac{3013\frac{3}{4}}{780}$. He now bisects the angle BAD getting the line AE and proceeds in exactly the same way to get an upper bound for $\frac{AB}{BE}$. He bisects two more times getting the line AG. Using the same technique he gets the estimate

$\frac{AB}{BG} < \frac{2017\frac{1}{4}}{66}$. So $\frac{GB}{AB} > \frac{66}{2017\frac{1}{4}}$. Since GB is a side of a regular inscribed polygon with 96 sides we see that the ratio of the perimeter of the polygon to the radius of the circle is greater than $\frac{66\times 96}{2017\frac{1}{4}} > 3\frac{10}{71}$.

There are many theories as to how Archimedes found his accurate upper and lower bound for $\sqrt{3}$ one that is very convincing can be found in A.Weil, Number Theory, Birkhäuser, Boston, 1984. He suggests that Archimedes was applying the formula

$$(5x+9y)^2 - 3(3x+5y)^2 = -2(x^2 - 3y^3).$$

Then according to Weil he started with $x = 1$ and $y = 0$ and since

$$5^2 - 3\times 3^2 = -2.$$

We have

$$\left(\frac{5}{3}\right)^2 = 3 - \frac{2}{9}.$$

The iteration involves two parts

$$(x,y) \to \left(\frac{5x+9y}{2}, \frac{3x+5y}{2}\right)$$

then

$$(x,y) \to (5x+9y, 3x+5y).$$

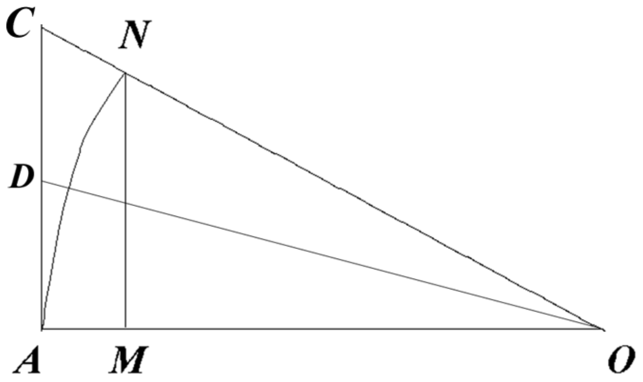
The first iteration ($x = 5, y = 3$) yields $26^2 - 3(15)^2 = 1$, that is $\left(\frac{26}{15}\right)^2 = 3 + \frac{1}{15^2}$. In the second ($x = 26, y = 15$) one has $5\times 26 + 9\times 15 = 265$, $3\times 26 + 5\times 15 = 153$ and $(265)^2 - 3(153)^2 = -2$ thus $\left(\frac{265}{153}\right)^2 = 3 - \frac{2}{(153)^2}$. This gives Archimedes lower bound. The upper bound is obtained by putting $x = 265$ and $y = 153$ into the first formula. Getting the pair

$$\frac{5\cdot 265 + 9\cdot 153}{2} = 1351, \frac{3\cdot 265 + 5\cdot 153}{2} = 780$$

the upper bound used by Archimedes.

2.4.4 Exercises.

1. Consider the diagram (the curve AN should be an arc of the circle of radius OA)



with OD the bisector of the angle AOC show that $\frac{OC}{OA} = \frac{CD}{DA}$. (Hint: Use trigonometry. Let θ be the angle AOC. Then we are asked to show that

$$\frac{\tan\theta - \tan\frac{\theta}{2}}{\tan\frac{\theta}{2}} = \frac{1}{\cos\theta}.$$

Do this by using the usual trigonometry identities $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$ and $\cos\theta = (\cos\frac{\theta}{2})^2 - (\sin\frac{\theta}{2})^2$.)

2. In the calculation of the upper bound for π Archimedes replaced the better estimate $3 + \frac{667\frac{1}{2}}{4673\frac{1}{2}}$ by $3\frac{1}{7}$. Why do you think he did it?

3. The Babylonians preferred the upper bound $3\frac{1}{6}$ over $3\frac{1}{7}$ for π. Can you give a reason for this?

4. Do the indicated iterations at the end of this section for sharper and sharper approximations to $\sqrt{3}$. What would you do to get good approximations of $\sqrt{5}$?

2.4.5 The iteration in Archimedes calculation.

It is clear that Archimedes could have in principle continued his bisection procedure indefinitely. However, the calculations become more and more complicated and even powerful arithmetician (as Archimedes obviously was) would be stymied by the calculation after two more bisections. Furthermore, little would have been gained since his initial choices of square roots of 3 limited him to about 4 digits of accuracy. We now live in an age of cheap high speed calculation power and can therefore implement many more iterations of Archimedes method. Let is first abstract the iteration that Archimedes does 4 times for the upper bound and 4 times for the lower.

We will use the diagram of Exercise 1. above we take the angle AOC to be such that $2m$ times it makes exactly one rotation. Let θ denote that angle. Then $2AC$ is a side of the regular m-gon circumscribed on the circle of radius OA and $2NM$ is the side of the regular m-gon inscribed in the same circle. Thus the circumference of the circumscribes m-gon is $2mAC$ and the circumference of the inscribed is $2mNM$. We observe that $\frac{AC}{OA} = \tan\theta$ and $\frac{NM}{OA} = \sin\theta$. Thus the circumscribed circumference divided by the diameter $(2OA)$ is $a = m\tan\theta$ and the circumference of the inscribed divided by the diameter is $b = m\sin\theta$. We now bisect the angle AOC getting AOD. Then using the same argument we find that the circumference of the circumscribed $2m$-gon divided by the diameter is $a' = 2m\tan\frac{\theta}{2}$ and the corresponding ratio for the inscribed $2m$-gon is $b' = 2m\sin\frac{\theta}{2}$. The key point is

$$a' = \frac{2ab}{a+b},$$

$$b' = \sqrt{a'b}.$$

Let us check this with standard trigonometry. We will use $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$ and $\cos\theta = (\cos\frac{\theta}{2})^2 - (\sin\frac{\theta}{2})^2$. If we use the first identity we find that

$$a'b = \frac{2m\sin\frac{\theta}{2}(2m\sin\frac{\theta}{2}\cos\frac{\theta}{2})}{\cos\frac{\theta}{2}} = 4m^2(\sin\frac{\theta}{2})^2 = (b')^2.$$

This shows that the second identity is true. As for the first

$$\frac{2ab}{a+b} = \frac{2m^2\frac{(\sin\theta)^2}{\cos\theta}}{m(\frac{\sin\theta}{\cos\theta} + \sin\theta)} = \frac{2m\sin\theta}{1+\cos\theta} = \frac{4m\sin\frac{\theta}{2}\cos\frac{\theta}{2}}{1 + (\cos\frac{\theta}{2})^2 - (\sin\frac{\theta}{2})^2}.$$

We now use the fact that $1 = (\cos\frac{\theta}{2})^2 + (\sin\frac{\theta}{2})^2$ so the denominator of the last expression is $2(\cos\frac{\theta}{2})^2$. Substituting we find that the last expression is $\frac{4m\sin\frac{\theta}{2}\cos\frac{\theta}{2}}{2(\cos\frac{\theta}{2})^2} = 2m\tan\frac{\theta}{2}$.

We will now use these observations to set up the implied iteration in Archimedes. We start with $m = 6$ then Archimedes has shown that $\tan\theta = \frac{\sqrt{3}}{3}$ and $\sin\theta = \frac{1}{2}$. Thus the corresponding ratios (which we denote by a_0 and b_0) are $a_0 = 2\sqrt{3}$ and $b_0 = 3$. If we do the first bisection then we have a_1 and b_1 with

$$a_1 = \frac{2a_0b_0}{a_0+b_0}$$

and

$$b_1 = \sqrt{a_1b_0}.$$

In general we have after $n+1$ bisections

$$a_{n+1} = \frac{2a_nb_n}{a_n+b_n}$$

and

$$b_{n+1} = \sqrt{a_{n+1}b_n}.$$

Archimedes is using the upper and lower bounds $b_4 < \pi < a_4$ and extremely clever choices of approximate square roots. If you do the calculation using a computer you find that $a_4 = 3.14271...$ and $b_4 = 3.14103...$ whereas Archimedes estimates are $\frac{22}{7} = 3.142857...$ and $3\frac{10}{71} = 3.140845...$ The estimates of Archimedes are therefore truly remarkable. We note that $a_{10} = 3.141592930...$ and $b_{10} = 3.141592519...$ Thus 4 iterations gives 2 decimal place accuracy and 10 gives 6 decimal place accuracy. One finds that after 16 iterations a_{16} and b_{16} agree to 8 decimal places. This predicts that $2k$ iterations should give an accuracy of k decimal places (indeed a calculation shows that one has 50 digit accuracy after 100 iterations). This can be (essentially) proved as follows. We note that

$$\begin{aligned} a_{n+1} - b_{n+1} &= \frac{2a_nb_n}{a_n+b_n} - \sqrt{\frac{2a_nb_n^2}{a_n+b_n}} = \\ & \frac{\sqrt{2a_n}b_n}{(a_n+b_n)(\sqrt{2a_n}+\sqrt{a_n+b_n})}(a_n-b_n). \end{aligned}$$

One can check that $\frac{\sqrt{2a_n}b_n}{(a_n+b_n)(\sqrt{2a_n}+\sqrt{a_n+b_n})} \leq \frac{1}{2+\sqrt{2}} < \frac{1}{3}$. To see this we first note that for all $n \geq 0$

$$b_n < a_n$$

indeed if $n = 0$ this is the assertion that $2\sqrt{3} > 3$. Assuming this for n we note that the iteration implies that

$$b_n = \frac{a_{n+1}(a_n+b_n)}{2a_n}.$$

Thus

$$b_{n+1} = \sqrt{\frac{a_{n+1}^2(a_n+b_n)}{2a_n}} = a_{n+1}\sqrt{\frac{a_n+b_n}{2a_n}} < a_{n+1}$$

since $a_n > b_n$. Thus the principle of mathematical induction proves the assertion. In

$$\frac{\sqrt{2a_n}b_n}{(a_n+b_n)(\sqrt{2a_n}+\sqrt{a_n+b_n})}$$

we divide the numerator and denominator by $\sqrt{2a_n}b_n$ and get

$$\frac{1}{\left(1+\frac{a_n}{b_n}\right)\left(1+\sqrt{\frac{a_n+b_n}{2a_n}}\right)}.$$

We now observe that $\frac{a_n}{b_n} > 1$ and $\sqrt{\frac{a_n+b_n}{2a_n}} > \frac{1}{\sqrt{2}}$. Thus the expression is less than $\frac{1}{2(1+\frac{1}{\sqrt{1}})} = \frac{1}{2+\sqrt{2}}$.

There is something very odd about this recursion that is that the expression for b_{n+1}involves a_{n+1}. Consider the following change in the recursion:

$$\begin{aligned} a_{n+1} &= \frac{2a_n b_n}{a_n+b_n} \\ b_{n+1} &= \sqrt{a_n b_n}. \end{aligned}$$

One checks that with this recurrence starting with the same initial values we have agreement between a_5 and b_5 to 50 decimal places. There is only one problem with replacing Archimedes iteration with this one. It converges to 3.219546022.... Which could be an interesting number but it is not π. What is it? The next number will unravel this mystery and lead to a method of determining π to very high orders of accuracy. Before we go on to these developments we will make one general observation about the above iteration we first note that. The iteration implies that we have

$$b_{n+1} - a_{n+1} = \frac{\sqrt{a_n b_n}}{(a_n+b_n)}(a_n + b_n - 2\sqrt{a_n b_n}) =$$

$$\frac{\sqrt{a_n b_n}}{(a_n+b_n)}\left(\sqrt{b_n} - \sqrt{a_n}\right)^2.$$

This implies that if $n > 1$ then $b_n > a_n$. We note that $\left(\sqrt{b_n} - \sqrt{a_n}\right)\left(\sqrt{b_n} + \sqrt{a_n}\right) = b_n - a_n$. We therefore have

$$b_{n+1} - a_{n+1} = \frac{\sqrt{a_n b_n}}{(a_n+b_n)\left(\sqrt{b_n} + \sqrt{a_n}\right)^2}(b_n - a_n)^2.$$

We estimate the expression $\frac{\sqrt{a_n b_n}}{(a_n+b_n)\left(\sqrt{b_n}+\sqrt{a_n}\right)^2}$. For simplicity we assume that $a_0 \geq 1$ and $b_0 \geq 1$ we assert that $a_n \geq 1$ and $b_n \geq 1$ for all n. If $n = 0$ this is our assumption. If we assume this assertion for n then $a_n b_n \geq a_n$ and $a_n b_n \geq b_n$ so $2a_n b_n > a_n + b_n$ hence $a_{n+1} \geq 1$ also $a_n b_n \geq 1$ implies that $\sqrt{a_n b_n} \geq 1$ so $b_{n+1} \geq 1$. We have already seen that $a_n + b_n - 2\sqrt{a_n b_n} = \left(\sqrt{b_n} - \sqrt{a_n}\right)^2$ so $a_n + b_n \geq 2\sqrt{a_n b_n}$. Thus we have

$$\frac{\sqrt{a_n b_n}}{a_n + b_n} \leq \frac{1}{2}$$

and since $a_n \geq 1$ and $b_n \geq 1$ we see that $\left(\sqrt{b_n} + \sqrt{a_n}\right)^2 \geq (1+1)^2 = 4$. Thus

$$\frac{\sqrt{a_n b_n}}{(a_n+b_n)\left(\sqrt{b_n} + \sqrt{a_n}\right)^2} \leq \frac{1}{8}.$$

This implies that in the modified iteration we have

$$b_{n+1} - a_{n+1} \leq \frac{1}{8}(b_n - a_n)^2$$

for $n \geq 0$ if a_0 and b_0 are both at least 1. Actually one has a similar estimate if we only assume that a_0 and b_0 are bigger than 0 (we will see why in the next section when we relate this iteration to the arithmetic-geometric mean iteration.. This accounts for the rapid convergence. Starting with $a_0 = 2\sqrt{3}$, $b_0 = 3$. Then $b_1 - a_1 \leq \frac{1}{8}(a_0 - b_0)^2 = 0.0269238...$ Now

$$b_2 - a_2 \leq \frac{1}{8}(b_1 - a_1)^2 \leq \frac{1}{8^3}(b_0 - a_0)^4 = 0.00009063...,$$

$$b_3 - a_3 \leq \frac{1}{8}(b_2 - a_2)^2 = \frac{1}{8^7}(b_0 - a_0)^8, ..., b_n - a_n \leq \frac{(b_0 - a_0)^{2^n}}{8^{2n+1}}, ...$$

2.4.6 The arithmetic-geometric mean iteration of Gauss.

Recall the new iteration of the previous subsection:

$$\begin{aligned} a_{n+1} &= \frac{2a_n b_n}{a_n + b_n} \\ b_{n+1} &= \sqrt{a_n b_n}. \end{aligned}$$

With a_0, b_0 positive real numbers. If we write $a_n = \frac{1}{u_n}$ and $b_n = \frac{1}{v_n}$ then we have the recursion

$$\begin{aligned} u_{n+1} &= \frac{u_n + v_n}{2} \\ v_{n+1} &= \sqrt{u_n v_n}. \end{aligned}$$

The first is the *arithmetic mean* of u_n and v_n and the second is their *geometric mean*. If u and v are positive then their mean or arithmetic mean is their average $a(u,v) = \frac{u+v}{2}$ their geometric mean or multiplicative average is $m(u,v) = (uv)^{\frac{1}{2}}$. We note that $a(u,v)^2 - m(u,v)^2 = \frac{(u-v)^2}{4} \geq 0$. Thus if we start the iteration with $u_0, v_0 \geq 0$ then $u_n \geq v_n$ for all $n \geq 1$. This iteration was discovered independently by J.L.Lagrange (1736-1813) and C.F.Gauss (1777-1855). Lagrange alluded to it in 1785 and Gauss studied it in 1790 (when he was about 14). We attach the name of Gauss since he did the most profound work on it in particular answering the question we asked at the end of the last section. The iteration is called the *AGM* On May 30, 1799 Gauss wrote (in his diary) that if we start the iteration with $u_0 = 1$ and $v_0 = \sqrt{2}$ then $\frac{1}{u_n}$and $\frac{1}{v_n}$ are equal to

$$\frac{2}{\pi}\int_0^1 \frac{dt}{\sqrt{1-t^4}}$$

to at least 11 decimal places for n large. Notice that he is predicting the value of the original variant of the Archimedean iteration. He was absolutely certain that the limit of the sequences was in fact this number. In his diary he said that should this be true then it “will surely open a whole new field of analysis”. The area of analysis that was opened is the theory of elliptic functions which is still one of the most important areas of mathematics that has permeated every aspect of the science and about which we shall hear much more later. Before giving Gauss's solution to the general problem we will explain a possible reason why he might believe that the limit in the above case might be given by an integral of the above sort. We first return to the Archimedean iteration of the previous section

$$\begin{aligned} a_{n+1} &= \frac{2a_n b_n}{a_n + b_n} \\ b_{n+1} &= \sqrt{a_{n+1} b_n} \end{aligned}$$

with $a_0 = \tan\theta$, $b_0 = \sin\theta$ and $0 < \theta < \frac{\pi}{2}$. Then as above we see that there is a number L such that $b_n < L < a_n$ and that $a_n - b_n$ can be made as small as we wish by increasing n. The amazing fact is that the number L is θ. Thus for example if we start with $a_0 = 1$, $b_0 = \frac{1}{\sqrt{2}}$ then $\theta = \frac{\pi}{4}$. This could have lead him to look at his iteration multiplied by $\frac{\pi}{4}$. He was no doubt certain that integrals of the type of his projected formula for his limit could not be calculated using elementary methods (e.g. modern Freshman or Sophomore Calculus). The integral

$$\int_0^1 \frac{dt}{\sqrt{1-t^4}}$$

is one of the simplest of the type that he would have studied. He therefore would have known that it was about 1.311028777. From this and his no doubt very accurate approximation to $\frac{\pi^2}{8}$ he could have easily come up with the approximation in his diary entry of 1799.

Gauss later derived a formula for the limit of the AGM which can be found in volume 3 of his collected works. The solution is given in terms of a completed elliptic integral. We will just quote the formula (a very nice discussion can be found in Borwein and Borwein, Pi and the AGM. Consider the AGM then since $\frac{xa+xb}{2} = x\left(\frac{a+b}{2}\right)$ and $\sqrt{(xa)(xb)} = x\sqrt{ab}$ for $x, a, b > 0$ we see that if we denote be $M(a,b)$ the limit of the AGM with $u_0 = a, v_0 = b$. Then if $a, b > 0$, $M(a,b) = aM(1, \frac{b}{a})$. It is therefore enough to calculate $M(1,x)$ for $x > 0$. Here is the formula

$$\frac{1}{M(1,x)} = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \frac{d\theta}{\sqrt{1-(1-x^2)\sin^2\theta}}.$$

2.4.7 Exercises.

1. Consider the following iteration

$$\begin{aligned} a_{n+1} &= \frac{a_n + b_n}{2} \\ b_{n+1} &= \frac{2a_n b_n}{a_n + b_n}. \end{aligned}$$

With $a_0 > 0, b_0 > 0$. Show by induction that $a_n b_n = a_0 b_0$. Also that

$$a_{n+1} - b_{n+1} = \frac{(a_n - b_n)^2}{2(a_n + b_n)}.$$

Use these observations to derive a very fast method of calculating square roots. (Note that if we start with $a_0 = \frac{3}{2}$, $b_0 = 2$ then $a_2 = \frac{97}{56}, b_2 = \frac{168}{97}$ and $a_3 = \frac{18817}{10864}, b_3 = \frac{32592}{18817}$, further $(b_3)^2 = 2.999999992...$

2. Use the method in section 2.4.5 of the derivation of the Archimedean iteration to show that if $\theta = \frac{\pi}{m}$ then the Archimedean iteration (the main iteration in 2.4.5) starting with $a_0 = \tan\theta$, and $b_0 = \sin\theta$ eyelids θ in the limit.

3. Make the appropriate change of variables to show that the value of $M(1, \sqrt{2})$ using Gauss's general formula agrees with the one he predicted in 1799.

2.4.8 A short history of calculations of π.

As we have seen, Archimedes is the author of the famous approximation $\frac{22}{7}$ for π. We have also seen that most ancient peoples who were aware of π used 3. The Babylonians used the somewhat better approximation $3\frac{1}{6}$. We will end this short history with a method of approximation based on the AGM.

After Archimedes the iteration described above was used to find approximations to π until the 17th century.perhaps the best usage (and perhaps the last) was by Ludolph van Ceulen (1540-1610) who used the method to calculate 34 digits of π. This method converges relatively slowly so the computational overhead overwhelms hand computation. It wasn't until the advent of calculus that more precise approximations were found using more rapidly converging sequences. Until the middle of the twentieth century the approaches involved clever uses of a formula attributed to James Gregory (1638-1675) which says that

$$\arctan(x) = x - \frac{x^2}{3} + \frac{x^4}{5} - \frac{x^6}{7} + ...$$

If, in this series, we set $x = 1$ then we have

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + ...$$

Edmond Halley (1656-1743) used $x = \sqrt{1/3}$ in Gregory's series to produce the series

$$\frac{\pi}{6} = \frac{1}{\sqrt{3}}(1 - \frac{1}{3 \cdot 3} + \frac{1}{3^2 5} - \frac{1}{3^3 7} + ...).$$

He used this to find π to 71 decimal places to do this he needed to sum at least 143 terms of the series. The approximation is $\frac{p}{q\sqrt{3}}$ with

$p =$ 29757519334033066607627810378515807324571797252183413378517664256040092164338566715216074032725294059375304662800878489724243784035047966009731713992494875785074158410997535604864857105476 48

and

$q =$ 5468726754975023858173190008331026443083349550027969750
06063504744927456329014146000945985504325020793071970588029
48449190349218434866194124401527196795946520854577134466195
5929457343724625

This is an amazing achievement using hand calculation.

Later variants of these methods are related to the formula of John Machin (1680-1752). He observed that

$$\frac{\pi}{4} = 4\arctan(\frac{1}{5}) - \arctan(\frac{1}{239})$$

the point here is that the first term is easily calculated and the second is an alternation of terms that become very small rapidly. Machin used his formula to find 100 digits of π.

In 1961 using an IBM mainframe D. Shanks and J.W.Wrench produced 100,000 digits of π using two 3 term variants of Machin's formula. (one to check the other). Using the same method one million digits were computed in 1973 by Guillard and Bouyer. Further precision has for the most part been based on the AGM. As of 1999 the record is held by Kanada, Takahashi 1999 with 206158430000 digits.

We include here an iterative scheme for calculating π that was discovered by Borwein and Borwein in the 1980's that is derived from the AGM. The iteration is as follows:

$$\begin{aligned}
x_{n+1} &= \frac{1}{2}\left(\frac{x_n+1}{\sqrt{x_n}}\right) \qquad n \geq 0 \\
y_{n+1} &= \frac{y_n x_n + 1}{\sqrt{x_n}(1+y_n)} \qquad n \geq 1 \\
\pi_n &= \pi_{n-1}\frac{x_n+1}{y_n+1} \qquad n \geq 1
\end{aligned}$$

with $x_0 = \sqrt{2}, \pi_0 = 2+\sqrt{2}, y_1 = \sqrt[4]{2}$. One can show that if $n \geq 2$

$$10^{-2^{n+1}} \geq \pi_n - \pi \geq 0.$$

This says that after 10 iterations we have π to the accuracy of over 2000 decimal digits. After 20 we heve over 2000000 digits.

2.4.9 Exercise.

1. Use a computer algebra system to check that the iteration does indeed give the asserted accuracy for $n = 2, 3, 4, 5$ (asserted $8, 16, 32, 64$ digits). Devise an algorithm combining 2.4.7 Exercise 1 with the above iteration to get a high precision algorithm for π with no square roots.

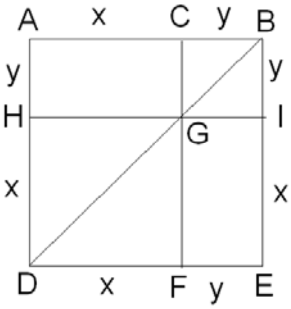
3 The emergence of algebra

As we saw in chapter one the ancient Babylonians and Egyptians had an understanding of much of what we would now call Algebra. For the most part their algebra comes down to us in the form of problems that are not very different from those that are assigned to modern students. The ancients certainly had an understanding of how to solve linear and quadratic equations in one variable. The Babylonians with the use of copious tables could solve some cubic equations. But they were hampered in their lack of two basic formalisms that we take for granted. The first is that they had no concept of negative numbers and they had no notation such as our modern algebra which allowed them to handle an unknown quantity as if it were a number. There was still a basic distinction between the role of numbers for counting and numbers for measurement. As we shall see, the final synthesis involves the identification of the two notions of number.

3.1 Algebra in Euclid's Elements.

In Euclid's elements Book II can be considered to be devoted to algebra. For example, Proposition 4 book II says:

If a straight line be cut at random, the square on the whole is equal to the squares on the segments and twice the rectangle contained by the segments.



We will not go through Euclid's proof here (which is surprisingly long). We will just point out that what it says is that the square $ADEB$ is made up of the two squares $CBIG$, $DFGH$ and the two equal rectangles $ACGH$ and $FEIG$. In more modern notation the side of the big square is $AB = x + y$. The side of the square $CBIG$ is y that of $DFGH$ is x and the two adjacent sides of the two rectangles are x, y. Thus the content of the Proposition is

$$(x + y)^2 = x^2 + 2xy + y^2.$$

We will see that until the time of Descartes, the part of mathematics that we consider to be algebra was consistently phrased in geometric terms. In Euclid's number theory a number was a concatenation of unit intervals. He only considers whole numbers. In his geometric algebra (Book II) he considers lengths and areas but gives no direct relationship with the concept of number which comes later (Book VII). Thus intervals, squares and rectangles, cubes are dealt with as if they are what we consider to be numbers. The addition and subtraction meant putting together figures as in the one above. The amazing aspect of all of this is that within these constraints mathematicians were able to do serious work in algebra such as solving a polynomial of degree3 or 4. The constraints were broken by the seventeenth century French mathematicians.

3.2 The Arabian notation.

In chapter 1 we studied the methods that were used by several early cultures to represent numbers. We also looked at our own decimal system. This positional system was used by the peoples of the middle east and comes to us under the name Arabic notation. This notation when it appeared in Europe was very similar to our modern notation (however it is likely that it had its genesis in India). One of the most important and earliest western advocates of this system was Leonardo of Pisa (alias Fibonacci) who used the system in his book *Liber Abaci* (published in 1228). In this book he used the arabic notation for everything but fractions. For fractions he used sexagesimal, Egyptian fractions and common fractions (that is a/b in lowest terms). He preferred the latter two types. We have seen that he devised an algorithm to convert common fractions to Egyptian fractions. We also observed that he gave a complete characterization of Pythagorean triples.

3.2.1 The completion of the characterization of Pythagorean triples.

The following argument involves the understanding of squares of integers. Fibonacci was so enamored of squares that he wrote a book *Liber Quadratorum* (Book of Squares) which contained the proof of the following theorem (see also Euclid Book X, Lemmas 1,2 before Proposition 29):

If a, b, c are positive integers such that $a^2+b^2 = c^2$ then one of a and b, say, b must be even and there exist numbers m, n, x such that $a = x(m^2 - n^2), b = 2xmn, c = x(m^2 + n^2)$. If a, b are relatively prime we can take $x = 1$.

To prove this assertion we first show that one of a or b must be even. Suppose not. Then $a = 2r+1, b = 2s+1$ and so $c^2 = a^2+b^2 = 4r^2+4r+1+4s^2+4s+1 = 4(r^2+r+s^2+s)+2$. We can thus conclude that c^2 is even. This can only be so if c is even. But then $c = 2t$. We conclude that $4t^2 = 4(r^2+r+s^2+s)+2$. This leads to the conclusion that 4 divides 2. We can thus assume that b is even. To complete the hard part of the argument we first observe that the last assertion implies the main assertion. Indeed, if x divides a and b then x^2 divides c^2 so x divides c. (You should be starting to see why the book had its title.). Thus if a, b, c is a Pythagorian triple and if x is the greatest common divisor of a

and b then $\frac{a}{x}, \frac{b}{x}, \frac{c}{x}$ is a Pythagorian triple. So we are left with showing the last assertion. We thus may assume a and b are relatively prime and that b is even.

Since $a^2 + b^2 = c^2$ it follows that $c^2 - a^2 = b^2$. So $(c-a)(c+a) = b^2$. We notice that if c were even then a must be even. But then a and b would have 2 as a common factor. Thus a and c are odd. If y is odd and divides both $c+a$ and $c-a$ then y divides their sum and difference which are $2c$ and $2a$. But then x^2 divides b^2. So x divides b which is contrary to our assumption. Thus if p is an odd prime so that p^r divides b then p divides exactly one of $c+a$ and $c-a$ thus p^{2r} divides one of the factors. Hence if $b = 2^t p_1^{r_1} \cdots p_u^{r_u}$ is a prime factorization of b then we can reorder the indices so that $c + a = 2^v p_1^{2r_1} \cdots p_s^{2r_s}$ and $c - a = 2^w p_{s+1}^{2r_{s+1}} \cdots p_u^{2r_u}$ with $v + w = t$. If v and w were both bigger than 1 then 4 would divide both $2c$ and $2a$. Since both are at least one we see that one of w and v must be one. This implies that the other must be odd. If $v = 1$ then $c+a = 2m^2$ and $c-a = 2n^2$. It is clear that if $w = 1$ we come to the same conclusion. Thus $b^2 = 4m^2n^2$ so $b = 2mn$. Also $2c = (c+a)+(c-a) = 2(m^2+n^2)$ and $2a = (c+a) - (c-a) = 2(m^2 - n^2)$. So a, b, c are of the desired form.

3.2.2 Exercises.

1. Observe that if $m = 2, n = 1$ then $m^2 - n^2 = 3$, $2mn = 4$, and $m^2 + n^2 = 5$. If $m = 3$, $n = 1$ then $m^2 - n^2 = 8$, $2mn = 6$ and $m^2 + n^2 = 10$. Thus aside from the factor of 2 and the order the two give the same Pythagorian triple. Show that if m, n are relatively prime then the greatest common divisor of any pair of the Pythagorian triple is either 1 or 2.

2. Show that if x, y are rational numbers such that $x^2 + y^2 = 1$ then there exist integers m, n such that either

$$(x, y) = (\frac{m^2 - n^2}{m^2 + n^2}, \frac{2mn}{m^2 + n^2})$$

or

$$(x, y) = (\frac{2mn}{m^2 + n^2}, \frac{m^2 - n^2}{m^2 + n^2}).$$

(Hint: $y = \pm\sqrt{1 - x^2}$. If $x = \frac{a}{c}$ in lowest terms then $1 - x^2 = \frac{c^2 - a^2}{c^2}$. If the square root is rational then $c^2 - a^2$ must be the square of an integer b. Thus $a^2 + b^2 = c^2$.)

3. What is the overlap between the sets described by the two formulas in problem 2?

4. If we divide the numerators and denominators of the first expression in problem 2 by n^2 and write $t = \frac{m}{n}$ then we have

$$(x.y) = (\frac{1 - t^2}{1 + t^2}, \frac{2t}{1 + t^2}).$$

Show that the only pair of real numbers (x, y) not covered by a value of t is $(-1, 0)$. This is called the *rational parametrization of the circle.*

5. Look at Euclid Book X. Lemmas 1,2 before Proposition 29 and write out what the assertions mean algebraically.

3.2.3 Polynomials of higher degree.

Not as well known is the fact that Fibonacci studied cubic equations. Studying algebraic equations was quite difficult in his time due to a lack of appropriate notation and since the algebra of Fibonacci was still the geometric algebra of Euclid. Furthermore, cube roots were not constructed in the plane geometry so they were somewhat more mysterious. In his book the *Flos* (1225) he studied some cubic equations in particular

$$x^3 + 2x^2 + 10x = 20.$$

He proved that this equation has no rational roots and even no roots of the form $u + \sqrt{v}$ with u and v rational. He also gave an approximate solution to the equation in sexagesimal (1;22,7,42,33,4,40).

The Middle Eastern mathematicians had methods of calculating roots of polynomials to arbitrary precision. Most notable is the work of Abul Kamil (850-930). Also Omar Khayyam (1050-1130) had interpretations of roots of certain cubics as intersections of conic sections.

One reason for the slow progress in general methods of solution of polynomial equations was the lack of good notation (which persisted into the seventeenth century) and a lack of the ability to manipulate unknowns and indeterminates. For example, the unknown quantity x made sense to them (even to the ancients) as "a quantity" one could also say a "cube" a "side" and a "face" none of which are known. Then a description of a cubic equation could be given as a cube added to 2 times a face added to 6 times as side is equal to 12. We would write this as

$$x^3 + 2x^2 + 6x = 12.$$

The mathematicians developed clever short hand notations for such expressions. However, they did not go to the next stage and replace the $2, 6, 12$ by indeterminates a, b, c thus getting

$$x^3 + ax^2 + bx = c.$$

Rather, they dealt with the specific equation with explicit coefficients and used techniques that could work with many other coefficients. We have seen this approach in Euclid. It persisted into the seventeenth century and the work of Viéte and Descartes.

3.2.4 Exercises.

1. Show that there are no rational solutions to the equation

$$x^3 + 2x^2 + 10x = 20.$$

(Hint: If $x = \frac{a}{b}$ in lowest terms then

$$a^3 + 2a^2b + 10ab^2 = 20b^3.$$

Thus every prime divisor of b divides a. Hence $b = 1$. Now conclude a divides 20 and check all of the cases.)

2. Convert Fibonacci's approximation to decimal and check that it is a good approximation.

3.3 The solution of the cubic and quartic

3.3.1 The Tartaglia, Cardano approach to the cubic.

In spite of the fact that modern algebraic notation did not exist in the sixteenth century the general solution to the cubic (degree 3) and to the quartic (degree 4) was deduced by the Italian mathematicians Niccolo Tartaglia (1500-1577) and Gironomo Cardano (1501-1576) for the cubic and Ludovico Ferrari (1522-1565) for quartic. The history of that endeavor is not the most savory in the annals of mathematics and in fact it is almost certain that the solution to the cubic is in fact due to Scipione del Ferro (1465-1526) but unpublished. It seems that neither Tartaglia nor Cardano were morally as strong as they were mathematicians. It also seems that Tartaglia's role in the solution of the cubic was much more substantial than that of Cardano, although he seems to have been influenced by the rumors that a solution by Ferro existed.. We will leave these historical questions aside and just point out that there is an English translation of the *Ars Magna* published by the M.I.T Press (1968), translated by T.R.Witner, also the book *A History of Mathematics* by C.B.Boyer, *et al* has an interesting discussion of this history and further references. What seems to be well documented is that Tartaglia could solve a general cubic of the form

$$x^3 + ax = b.$$

It is quite conceivable that Cardano's contribution was to reduce the general cubic

$$x^3 + cx^2 + dx = e$$

to this form. In modern notation this is a fairly simple task. Set $x = y - u$. Then the equation says

$$y^3 - 3uy^2 + 3u^2y - u^3 + cy^2 - 2cuy + cu^2 + dy - du = e.$$

So if $u = \frac{c}{3}$ then the equation is given in y as

$$y^3 + ay = b$$

with $a = d - \frac{c^2}{3}$ and $b = \frac{dc}{3} + e - \frac{2c^3}{27}$. This step seems to be truly minor in our notation. But in the sixteenth century the methods used to derive such formulas were completely geometric. Recall that cubes had to be interpreted as volumes and squares as areas. Also the formulas used had to be given as geometric properties of areas of geometric figures. The final problem was that negative numbers were still not allowed and so there were many variants of the equations that needed to be analyzed. For example

$$x^3 + ax + c = 0$$

was not seen as the same

$$x^3 + ax = b$$

with $b = -c$. Similarly, the term ax might be on the right hand side of the equation. In addition to all of these complications, there was still no direct way of dealing with general quantities such as a and b above (the x was better understood). Thus rather than write the equation above Cardano would consider (say)

$$x^3 + 3x = 4.$$

He would than say: Let the cube plus 3 times the side equal 4. The 3 and the 4 would take the place of the a and the b. He would then go through an equivalent geometric discussion to the one below with the special values of a and b. With these provisos we will now derive a solution to the above reduced form of the cubic.

The critical idea is to write $x = u - v$. Then substituting in the equation we have

$$u^3 - 3u^2v + 3uv^2 - v^3 + a(u - v) = b.$$

That is

$$u^3 - 3uv(u - v) - v^3 + a(u - v) = b.$$

If we take $3uv = a$ then the equation becomes

$$u^3 - v^3 = b.$$

Now $v = \frac{a}{3u}$. So upon substitution we have

$$u^3 - (\frac{a}{3})^3 \frac{1}{u^3} = b.$$

Multiply through by u^3 and we have

$$u^6 - bu^3 - (\frac{a}{3})^3 = 0.$$

Apply the quadratic formula to solve for u^3 and get

$$u^3 = \frac{b \pm \sqrt{b^2 + 4(\frac{a}{3})^3}}{2} = \frac{b}{2} \pm \sqrt{(\frac{b}{2})^2 + (\frac{a}{3})^3}.$$

Notice that Cardano must choose the plus sign. Now $v^3 = u^3 - b$. Thus we have (at least as a possibility)

$$x = \sqrt[3]{\sqrt{(\frac{b}{2})^2 + (\frac{a}{3})^3} + \frac{b}{2}} - \sqrt[3]{\sqrt{(\frac{b}{2})^2 + (\frac{a}{3})^3} - \frac{b}{2}}.$$

This is one of Cardano's solutions (depending on various signs as we have pointed out). Notice that in the course of this development we have made choices. However, if we assume that $a > 0$ and that the only cube roots we can have are positive then we can reverse the steps

$$\sqrt[3]{\sqrt{(\frac{b}{2})^2+(\frac{a}{3})^3}+\frac{b}{2}}\sqrt[3]{\sqrt{(\frac{b}{2})^2+(\frac{a}{3})^3}-\frac{b}{2}}$$

$$= \sqrt[3]{(\frac{b}{2})^2+(\frac{a}{3})^3-(\frac{b}{2})^2} = \sqrt[2]{(\frac{a}{3})^3} = \frac{a}{3}.$$

Thus $x^3 + ax = \left(\sqrt[3]{\sqrt{(\frac{b}{2})^2+(\frac{a}{3})^3}+\frac{b}{2}}\right)^3 - \left(\sqrt[3]{\sqrt{(\frac{b}{2})^2+(\frac{a}{3})^3}-\frac{b}{2}}\right)^3 = b.$

3.3.2 Some examples.

First let us give some examples of Cardano's formula. Consider the equation

$$x^3 + x^2 + x = 14.$$

The first step is to eliminate the x^2. According to the recipe above, taking $c = 1$, $d = 1$, $e = 14$ we are "reduced" to

$$y^3 + ay = b$$

with $a = \frac{2}{3}$ and $b = \frac{385}{27}$. We can no plug into the formula

$$y = \sqrt[3]{\sqrt{\left(\frac{2}{9}\right)^3+\left(\frac{385}{54}\right)^2}+\frac{385}{54}} - \sqrt[3]{\sqrt{\left(\frac{2}{9}\right)^3+\left(\frac{385}{54}\right)^2}-\frac{385}{54}}.$$

Observe that this expression involves only square roots of positive numbers so at least it makes sense geometrically. If you do the calculation indicated you are looking at

$$y = \sqrt[3]{\frac{17}{18}\sqrt{57}+\frac{385}{54}} - \sqrt[3]{\frac{17}{18}\sqrt{57}-\frac{385}{54}}.$$

We now know that $y - \frac{1}{3}$ is a solution to our original equation. If you use a calculator and evaluate this expression numerically you will find that $y - \frac{1}{3}$ is approximately 2 and if you substitute 2 into the original equation you will find that 2 is indeed a solution. This indicates that there could be serious difficulties in the use of the elegant formula above. We will look at several other such difficulties in the exercises. Cardano and his contemporaries were much more worried about another problem. Which we will now describe.

First we must consider the form that was necessary for they must use to write the solution to

$$x^3 = ax + b.$$

without recourse to negative numbers. We would replace a by $-a$ and use the previous solution. Cardano did something equivalent using the substitution $x = u + v$. This gave rise to the solution

$$x = \sqrt[3]{\frac{b}{2} + \sqrt{(\frac{b}{2})^2 - (\frac{a}{3})^3}} + \sqrt[3]{\frac{b}{2} - \sqrt{(\frac{b}{2})^2 - (\frac{a}{3})^3}}.$$

If the term under the square root sign was non-negative then he had no trouble understanding the solution. However, if we consider (as Cardano did)

$$x^3 = 15x + 4.$$

The formula yields

$$\sqrt[3]{2 + \sqrt{-121}} + \sqrt[3]{2 - \sqrt{-121}}$$

which made no sense to Cardano. If you ask a mathematical software package to evaluate this expression numerically then it yields 4.0. Direct check shows that $x = 4$ is indeed a solution. The mathematics package would be hard pressed to see that this expression is *exactly* 4. (In Maple V version 4 the *simplify* operation doesn't yield 4. However, the *factor* operation does. Mathematica 4 (but not 3)actually returns 4 when it encounters this expression.)

We note that $x^3-15x-4 = (x-4)(x^2-4x+1)$. This implies that the equation has three distinct roots: 4 and the roots corresponding to the quadratic factor that involve square roots of positive numbers. Getting ahead of ourselves, we will see that if all three roots of a cubic are real and distinct Cardano's formula always involves a square root of a negative number (see exercise 3 below)

3.3.3 Exercises.

1. Show that the number $\sqrt[3]{\frac{17}{18}\sqrt{57} + \frac{385}{54}} - \sqrt[3]{\frac{17}{18}\sqrt{57} - \frac{385}{54}} - \frac{1}{3}$ is equal to 2 using Cardano's formula. (Show that the only real root of the corresponding equation is 2.)

2. Observe that if that if x is real then $\sqrt[3]{-x} = -\sqrt[3]{x}$. Use this to see that the choice made in the derivation of Cardano's formula didn't change the outcome.

3. Consider the equation $x^3 - 2x = 5$. Calculate the solution given by an appropriate variant of Cardano's formula. Next use a calculator or a computer to do Newton's iteration to derive an approximate solution (Newton actually did this calculation to 5 decimal places in 1669.) The Newton method is to guess a solution x_0. The iteration is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Here $f(x) = x^3 - 2x - 5$ and $f'(x) = 3x^2 - 2$. Thus if we start with the approximate root 2 then

$$x_1 = 2 + \frac{1}{10}.$$

Newton showed that x_2 was accurate to 5 decimal places.

4. Let $f(x) = (x-u)(x-v)(x-w)$ with u, v, w distinct and real. Assume that $u+v+w=0$. Show that $f(x) = x^3 - ax - b$ with $a = u^2 + uv + v^2$, $b = (uv)(u+v)$. Show that $(\frac{b}{2})^2 - (\frac{a}{3})^3$ is always negative. Thus if there are 3 real roots then the formula cannot be written directly in terms of real numbers.

3.3.4 The early attempts to explain the paradox.

This strange expression

$$\sqrt[3]{2+\sqrt{-121}} + \sqrt[3]{2-\sqrt{-121}}$$

that must be 4 was a thorn in the side of the remarkable achievement of solving the cubic (and relatively soon the quartic). The resolution of this paradox that the expression must be 4 but involves meaningless objects would not be fully resolved for about 400 years. As we shall see it goes to the heart of what we understand of numbers. Cardano had in earlier studies encountered other types of equations with solutions had a form

$$(u+\sqrt{-v}) + (u-\sqrt{-v})$$

and he knew that $2u$ was indeed a solution. Such numbers are now called *conjugate* complex numbers and we know that they do indeed add up to a number closer to the sense of Cardano and his contemporaries.

Rafael Bombelli (1526-1573) made a proposal the explain the paradox. He suggested the following "wild thought". Suppose the cube roots of a pair of conjugate numbers were conjugate? That is suppose we could write $\sqrt[3]{2+\sqrt{-121}} = u+\sqrt{-v}$ and $\sqrt[3]{2-\sqrt{-121}} = u-\sqrt{-v}$. Then the irksome sum would be $2u$. Since he "knew" that by all rights $2u$ should be 4 he chose $u=2$. He then computed $(2+\sqrt{-v})^3 = 8-6v+\sqrt{-(12-v)^2 v}$. Thus if Bombelli's wild thought is to work he must have $8-6v=2$. He was therefore forced to have $v=1$ and he found (probably to his own amazement) that $2+\sqrt{-121} = (2+\sqrt{-1})^3$ and $2-\sqrt{-121} = (2-\sqrt{-1})^3$. He now felt justified to plug his newfound cube roots into the Cardano formula and found that with his interpretation the Cardano solution was indeed equal to 4.

This brilliant analysis was no doubt very convincing at the time. However, from our perspective it leaves open more questions than it answers. However, before we begin to attempt to study the larger issues we will need to "bite the bullet" and understand what is meant by numbers. This will be begun in the next section when we discuss "analytic geometry". First we will give a short discussion of Ferrari's solution of the quartic and another related problem that arises from that result.

3.3.5 The solution of the quartic.

We first describe Ferrari's reduction of the solution of the quartic to the cubic. We are considering

$$x^4 + ux^3 + vx^2 + w = mx.$$

Cardano's technique for eliminating the square term in the cubic can be used to eliminate the cube. That is replace x by $x - \frac{u}{4}$. Then the equation is in the form

$$x^4 + ax^2 + b = cx.$$

The idea of Ferrari is to complete $x^4 + ax^2$ to a square by adding $\frac{a^2}{4}$ to both sides. We are thus looking at

$$(x^2 + \frac{a}{2})^2 = (\frac{a^2}{4} - b) + cx.$$

The critical step is to throw in another parameter (say) y in the left hand side and to observe that

$$\begin{aligned}(x^2 + \frac{a}{2} + y)^2 &= (x^2 + \frac{a}{2})^2 + 2(x^2 + \frac{a}{2})y + y^2 \\ &= (\frac{a^2}{4} - b) + cx + 2x^2y + ay + y^2.\end{aligned}$$

For the last equation we have substituted $(\frac{a^2}{4} - b) + cx$ for $(x^2 + \frac{a}{2})^2$. This term can be written in the form $Ax^2 + Bx + C$. With $A = 2y, B = c, C = (\frac{a^2}{4} - b) + ay + y^2$. We solve for y so that the quadratic equation has exactly one root. That so that $B^2 - 4AC = 0$ (i.e. we eliminate the ± term in the quadratic formula). Substituting the values of A, B, C we have

$$B^2 - 4AC = c^2 - 8y((\frac{a^2}{4} - b) + ay + y^2).$$

This is a cubic equation in y. Let u be a root of this equation (which we presumably can find using Cardano's formula). Then for this value we have

$$\begin{aligned}(x^2 + \frac{a}{2} + u)^2 &= (\frac{a^2}{4} - b) + cx + 2x^2u + au + u^2 \\ &= Ax^2 + Bx + C = A(x - (\frac{-B}{2A}))^2 \\ &= 2u(x + \frac{c}{4u})^2.\end{aligned}$$

That is

$$(x^2 + \frac{a}{2} + u)^2 = 2u(x + \frac{c}{4u})^2.$$

This says that to find a solution x we need only solve

$$x^2 + \frac{a}{2} + u = \sqrt{2u}(x + \frac{c}{4u}).$$

To do this we can apply the quadratic formula. The point of this is that in light of Cardano's formula we can write a solution to the quartic as an algebraic expression that involves arithmetic operations on square roots and cube roots of arithmetic operations on the coefficients of the equation. This is also true for the cubic and the quadratic formula does the same for degree 2. Of course, we must take into account the same provisos as we did for the cubic. When we study analytic functions of a complex variable we will come back to the sense in which Cardano's solution to the cubic (and thereby Ferrari's of the quadric) is actually a well defined solution.

In spite of these possible misgivings these results came at least 4000 years after the Babylonians understood how to solve the quadratic equation. The next natural problem was to find a solution of the quintic (fifth degree polynomial) in terms if arithmetic operations (addition, subtraction, multiplication, and division) and square roots, cube roots and fifth roots (radicals). The greatest mathematicians of that time and in fact for about the next 200 years could not find any clever method that would solve this problem. The answer to this problem was given by two of the most tragic cases in the history of mathematics. We will first discuss the solution of the problem for the quintic.

3.3.6 The quintic

The success of the Italian algebraists of the sixteenth century was extraordinary. The next step would be the quintic and then, of course, equations of higher degree. To the surprise of the mathematical community, there were no clever methods that they could find to reduce the solution of the quintic to that of the quartic, cubic and quadratic and extraction of fifth roots (or for that matter roots of any order). The prevailing idea had always been that one should be able to find the roots of any polynomial by doing arithmetic operations and extraction of roots. However, in 1799, Paolo Ruffini (1765-1822) published his two volume treatise *Teorie Generale delle Equazioni* in which he included an argument to show that there was no such method of solving the quintic. As happens in the history of mathematics, announced proofs of major new results are often incomplete or even wrong. Ruffini, in fact was on the right track but wrong in detail. One can imagine the scrutiny to which this treatise was subjected. The proof of the impossibility for the quintic was given a rigorous proof by Nicolas Abel (1802-1829) at the age of 19 (notice that he lived at most 27 years!). He published his proof in the form of a pamphlet at his own expense in 1824. Due to his limited funds, he had to keep the pamphlet brief and for that reason it was extraordinarily difficult to understand. He later proved a theorem that applied to all equations of degree 5 and higher. It states

If $n \geq 5$ then there is no formula for a solution involving arithmetic operations and extraction of roots on the coefficients of the equation

$$x^n + a_1x^{n-1} + a_2x^{n-2} + ... + a_n = 0.$$

This assertion is now known as the Abel-Ruffini theorem. A great deal of mathematics ocurred in the time that intervened between the work of Cardano, et. al. and the work of Abel. Most notably, the algebraic notation, which we now take for granted, was invented. Also the understanding and general usefulness if negative numbers was finally a standard part of mathematics.

Another major development in the interim was the invention of *complex numbers*. We will make a first (relatively geometric) attempt at explaining complex numbers in this chapter and will approach this concept more analytically in the next chapter. These numbers were used in so-called conjugate pairs in Bombelli's solution of the apparent paradox in the solution of the cubic. Within the system of complex numbers Carl Friedrich Gauss (1777-1855) proved the *fundamental theorem of algebra* which states

Within the complex numbers every equation

$$x^n + a_1 x^{n-1} + a_2 x^{n-2} + ... + a_n = 0$$

with $n > 0$ has a root.

Here we have two seemingly contradictory theorems. Abel asserts that there is no way of writing a formula (involving extraction of roots) for a solution if $n \geq 5$ and Gauss asserts that even so there is a solution. We will encounter such apparently paradoxical situations throughout our investigations. It should also be pointed out that Abel sent his pamphlet to Gauss who was acknowledged to be the most important mathematician of his time. Gauss was furious with the brevity of the work and made scathing remarks about it. Abel's article on this subject was eventually published in Crelle's Journal (founded in 1826) in the first issue which in addition contained 21 other articles by Abel. The theorem of Gauss will be studied in more detail in the later chapters (however, we will discuss an important special case later in this chapter). This theorem appeared in the thesis of Gauss and he went on to give numerous alternative proofs. The most notable aspect of the theorem is that it is not really a theorem in algebra. The reason for this is that, as we shall see, complex numbers (and in fact real numbers) are of a very different nature from integers and rational numbers. This difference is the basis of mathematical analysis (modern calculus).

Before we leave this subject in order to learn enough mathematics to discuss it in more depth, we should point out that the theorem of Abel only proves that there is no *general formula*. Obviously there are equations that we can solve by radicals. For example,

$$x^6 - 2 = 0.$$

It is thus reasonable to ask the question: What equations can be solved by radicals? Here we mean that even though we can't find a formula we can still write out a solution in terms of the coefficients of the equation using arithmetic (that is addition, multiplication and division) and extraction of roots (square roots, cube roots, fifth roots, etc.). Abel studied this problem also and laid the partial foundation of the mathematical theory of groups (terms like abelian groups are in the honor of Abel's work). However, the solution of this problem was completed by another teenager, Evariste Galois (1811-1832). Galois gave a complete criterion as to when an equation can be solved by radicals. But we are now well ahead of ourselves in this story. We now return to the situation at hand

and begin the development of a broader notion of number that would at least include the expressions described (recall Bombelli's analysis). In this broader formulation Gauss's fundamental theorem of algebra will hold. We will next need to introduce the theory of Galois (now called appropriately Galois Theory) which lays the basis of the theory of groups. The latter will be studied in later chapters. To begin the analysis of the first problem we must step back to the seventeenth century and study the work if Descartes.

3.4 Analytic Geometry.

René Descartes (1596-1650) is now mainly known for the notorious x, y axis and Cartesian coordinates (which we will see he never directly used) and for the quotation "I think therefore I am". Both are oversimplifications of what he actually did and what he actually meant. We will discuss his geometry which was an important beginning to what we now take for granted in algebra. The work that we will discuss is *The Geometry* which was published in 1637 as an addendum to his treatise *The Discourse on Method*. *The Geometry* consists of three books (we would probably call them sections). The first establishes a basis for the meaning of number in terms of geometry and establishes the notation that we still use today for polynomials with indeterminate coefficients. In the second he shows how one can use his algebraic methods to analyze plane figures in terms of polynomial equations. The third analyzes 3 and higher dimensions. It relates his notation with the earlier works of Cardano, et. al. and for example writes out Cardano's formula in exactly the same way we do.

There were several people whose work predated Descartes who understood the idea of independent variable and dependent variable. Nicole d'Oresme (1323-1382) actually did graphing of data much the way we do today (he did explicitly use what we call Cartesian coordinates thus a more accurate but cumbersome name might be Oresmian coordinates). We will have more to say about the work of this amazing man in the next chapter. Also, Francoise Viète (1540-1603) established our formalism of unknown quantities that we manipulate in the same way as if they were known numbers. Bombelli, in addition to his work on the mysteries of the cubic and the foundations of complex numbers wrote a treatise on algebra in which he also handled unknowns algebraically. Bombelli did not label his unknowns by letters, instead he invented a symbol for an unknown (not unlike our "frowning face" -(rotated ninety degrees). Fortunately that notation didn't catch on.

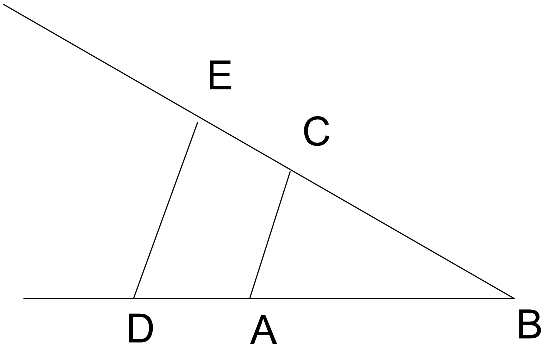
3.4.1 Descarte's notation and interpretation of numbers.

He begins his first book with the following assertion (we will use the translation given by David Eugene Smith and Marcia Latham):

> *Any problem in geometry can easily be reduced to such terms that a knowledge of the lengths of certain straight lines is sufficient for its construction. Just as arithmetic consists of only four or five operations, namely, addition,*

subtraction, multiplication and the extraction of roots, which may be considered a form of division, so in geometry, to find required lines it is merely necessary to add or subtract other lines; or else, taking one line which I shall call unity in order to relate it as closely as possible to numbers, and which can in general be chosen arbitrarily, and having given two other lines, to find a fourth which shall be to one of the given lines as the other is to unity (which is equivalent to multiplication);...

Before going on let us see what he means here. First he chooses a line segment which he calls unity and in the diagram below (which is essentially the same picture that occurs on page one in *The Geometry*) is denoted AB on the inclined line he measures BC (the first line) and on the line containing AB he measures BD. He then joins the points A and C. From D he draws the parallel line to AC which intersects the line containing BC at E. He then says that if we consider the ratios of BC and BD to AB then the ratio of BE to AB is the product of the corresponding ratios. Let us demonstrate the correctness of this assertion (Descartes feels no need to explain any more than what we have already said). The triangles ABC and DBE are similar. Thus the corresponding sides are all in the same proportion (Euclid, *Elements*, Book VI, Proposition 10). Thus $\frac{BE}{BD} = \frac{BC}{AB}$. If we think of BE as c times a unit, BC as a times a unit, BD as b times a unit and AB as 1 times a unit then the assertion is just that $a \times b = c$.



Descartes also had a method for doing division geometrically (we will give it as an exercise). We now come to an important point.

Often it is not necessary thus to draw lines on paper, but it is sufficient to designate each by a single letter. Thus to add the lines BD and GH, I call one a and the other b and write $a + b$. Then $a - b$ will indicate that the line b is subtracted from a; ab is the line a multiplied by b;...

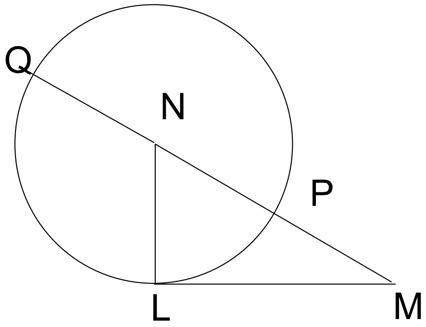
Here he is saying that $a+b$ is just the line corresponding to hooking together BD and GH on the same line. ab denotes the geometric operation of multiplication. The symbol a is a bit more abstract than a line since it corresponds to a line measured by a unit (shades of Euclid!). He writes division as we do and aa is a^2 and if this is multiplied by a then we have a^3, etc. The square root of $a^2 + b^2$ is denoted $\sqrt{a^2 + b^2}$. The cube root of $a^3 - b^3 + ab^2$ is written $\sqrt[3]{a^3 - b^3 + ab^2}$ and as he says similarly for other roots. Notice that so far he has only taken square roots of combinations of squares and cube roots of combinations of cubes or rectangular solids. Now comes the crux:

Here it should be observed that by a^2, b^3 and similar expressions, I ordinarily mean only simple lines, which however I name squares, cubes, etc., so that I may make use of the terms employed by algebra.

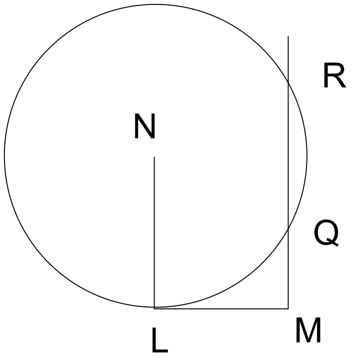
This means that even though we might be looking at a line segment of length 2 times the unit and then considering a cube that has one edge that segment we can think of the corresponding volume as a unitless number 8. This is obvious to us but it was not standard at the time. Further, it leads to our modern approach to numbers being produced by geometry. Descartes goes on to study the variants of the quadratic formula that are formed by changing the signs of the coefficients. Again his approach is quite modern and he writes such things as

$$x^2 = -ax + b^2.$$

To him negative numbers have a right to existence. However, since he still interprets the solution of the equation geometrically he looks three cases the above, $x^2 = ax + b^2$ and $x^2 = ax - b^2$. We show how Descartes handles the first two. Consider the following figure:



He takes LMN to be a right triangle with $LM = b$, $LN = \frac{a}{2}$. Now prolong MN to MQ a distance equal to NL. Then $x = MQ$ is the solution. If on the other hand we were considering the equation $x^2 = -ax + b$ then we would use the same figure from the point N we lay off NP on the line NM with the length of $NP = NL$. This time $x = PM$ is the answer. The last case is perhaps more interesting we are looking at $x^2 = ax - b^2$. For this we consider the following figure:



Here LM is of length b, LN is of length $\frac{a}{2}$ perpendicular to LM and the circle is of radius NL. The line through M is parallel to NL. There are three possibilities. The first is that the circle cuts the line through M in 2 points R, Q and both MQ and MR are solutions. The second is that the circle touches in one point, say Q, then MQ is the solution. The third is that $b > \frac{a}{2}$ that is the circle doesn't touch the line through M. Then he asserts that the equation has no solution. The main distinction between Descartes and his contemporaries is that his reason for the geometric constructions is the quadratic formula. So in the first example $x = MQ = QN + NM$. $QN = \frac{a}{2}$ and the Pythagorian theorem says that $NM = \sqrt{QN^2 + LM^2} = \sqrt{\frac{a^2}{4} + b^2}$ so $x = \frac{a}{2} + \sqrt{\frac{a^2}{4} + b^2}$. Which is the positive solution given by the quadratic formula. We will leave the other cases to the reader in the exercises.

The point here is that in Descartes formalism numbers and their units have been separated. Viète had allowed for the handling of unknowns as numbers but he still considered a product of two numbers to be an area, of three a volume. Thus $x^2 + 2x + 1$ makes sense to him as making an area that is a disjoint combination of a square of side x, a rectangle of side x and side 2 and a figure of area 1. For Descartes this "homogenization" is unnecessary.

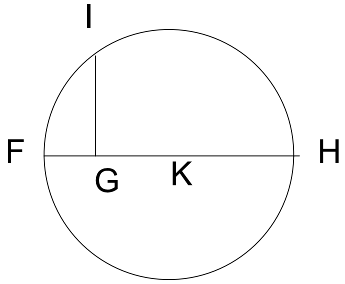
Even Descartes makes a distinction between positive solutions (actual) and negative (false). In the third case he has a third possibility no solution. In the *The Geometry* he has several results about counting actual solutions or converting false solutions into actual ones. Thus although he did not believe that negative numbers could be actual solutions to geometric problems he was aware of their existence in his algebraic formalism.

Book 2 of *The Geometry* is a study of curves in the plane. Although, the familiar x, y axes of analytic geometry do not appear explicitly in Descartes' work they are certainly implicit. We will come back to these ideas in the next chapter. In the next subsection we will consider his approach to what we now call analytic geometry.

3.4.2 Exercises.

1. Show how to divide using the same diagram as Descartes used to give a geometric interpretation of multiplication.

2. Consider the figure below. We quote Descartes: *If the square root of GH is desired, I add FG equal to unity; then bisecting FH at K, I describe the circle with radius FK with center K, and draw from G the perpendicular and extend it to I, and GI is the required root.* Show that this is indeed a geometric interpretation of the square root of GH.



3. Show that Descartes' geometric method does indeed describe the solutions to the quadratic equations described above.

3.4.3 Conics and beyond.

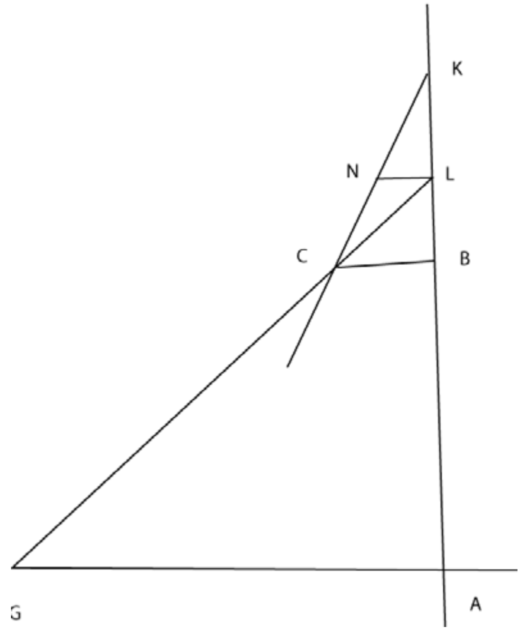
The second book of *The Geometry* contains the meat of the Cartesian method of algebraic geometry it has the title *On the nature of Curved lines.* The opening paragraph says:

The ancients were familiar with the fact that the problems of geometry may be divided into three classes, namely plane, solid, and linear problems. This is equivalent to saying that some problems require only circles and straight lines for their construction, while others require a conic section and still others require more complex curves. I am surprised, however, that they did not go further, and distinguish between different degrees of these more complex curves ...

The chapter ends with (the perhaps unwarranted) paragraph:

And so, I think I have omitted nothing essential to an understanding of curved lines.

Descartes begins by rejecting the study of certain curves such as spirals by saying that they "really belong only to mechanics". We will study one example from that chapter that first shows how configurations involving two lines (we will make this more precise) yield conic sections which can be described by quadratic equations and that if one if one allows a line to be a conic section then one has a cubic equation. We will also show how to use Descartes' method to derive an equation for an ellipse. We start with the following picture.

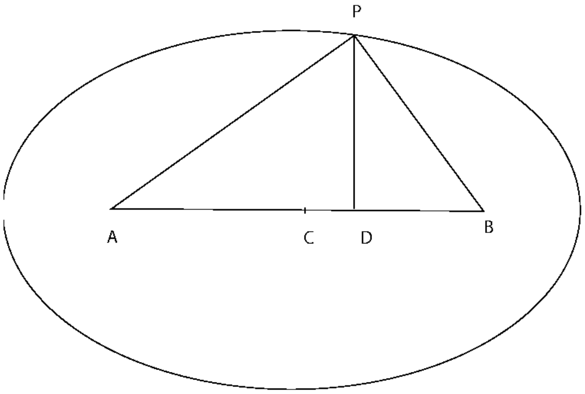


The angles at A, B, L are right angles. Descartes looks at the curve traced out as follows (this is just a paraphrase of what he actually says the algebra, however is the same as his). The points G and A are fixed. The dimensions of the figure KNL are fixed and the side KL is slid up and down the line AB. As it slides we look at the path that the point of intersection, C, of the line joining G to L and the line extending KN. He then observes that the two quantities BC and AB determine the point C. Since they are unknown he uses the notation y for BC and x for AB. (This is as close as he gets to the x, y axes.) The quantities that are known (or fixed) are AG which he calls a, KL which he calls b and LN which he calls c. He then uses similar triangles to observe that $\frac{KL}{NL} = \frac{BK}{BC.}$. Thus $\frac{b}{c} = \frac{BL+b}{y}$. That is, $BL = \frac{b}{c}y - b$. He uses similar triangles again to see that $\frac{GA}{AL} = \frac{BC}{BL}$. That is, $\frac{a}{x+BL} = \frac{y}{BL}$. This gives $aBL = y(x + BL)$. So $\frac{ab}{c}y - ab = yx + y(\frac{b}{c}y - b) = yx + \frac{b}{c}y^2 - by$. Multiplying through by $\frac{c}{b}$ we have $ay - ac = \frac{c}{b}yx + y^2 - cy$. We therefore have

$$y^2 = (a+c)y - \frac{c}{b}xy - ac.$$

Which Descartes observes is the equation of a hyperbola.

He then observes that if the figure to be slid were say a hyperbola then one would have gotten a higher order equation. Let us do one more example. We will follow his method to calculate an equation for the locus of points so that the sum of the distances from two fixed points is fixed (we now call the figure an ellipse). Consider the following diagram



P is the point on the ellipse. The two fixed points are A and B. C is the midpoint of the line segment AB joining A and B. The line PD is perpendicular to AB. We will call the constant value of the sum of the distances from P to A and to B, $2a$. We will also denote by c the length of the segment AC. Then $AP + PB = 2a$. We set $x = CD$ and $y = PD$. Then the Pythagorian theorem says that

$$AP^2 = (c+x)^2 + y^2,$$
$$PB^2 = (c-x)^2 + y^2.$$

Thus

$$\begin{aligned} AP^2 - PB^2 &= (c+x)^2 - (c-x)^2 \\ &= (c^2 + 2cx + x^2) - (c^2 - 2cx + x^2) = 4xc. \end{aligned}$$

Now

$$AP^2 - PB^2 = (AP - PB)(AP + PB) = 2a(AP - PB).$$

We therefore have

$$AP - PB = \frac{2xc}{a}.$$

Since $AP - PB + AP + PB = 2AP$ and by the above $AP - PB + AP + PB = \frac{2xc}{a} + 2a$ we have

$$AP = a + \frac{xc}{a}.$$

Similarly, $AP + PB - (AP - PB) = 2PB$. Thus

$$PB = a - \frac{xc}{a}.$$

We have $\left(a - \frac{xc}{a}\right)^2 = y^2 + (c-x)^2$. Hence $\left(a^2 - 2xc + \frac{x^2c^2}{a^2}\right) = y^2 + c^2 - 2xc + x^2$. This gives

$$x^2 + y^2 = a^2 - c^2 + \frac{c^2}{a^2}x^2.$$

It is convenient to write $b^2 = a^2 - c^2$. Then we have

$$x^2 + y^2 = b^2 + \frac{a^2 - b^2}{a^2}x^2 = b^2 + x^2 - \frac{b^2}{a^2}x^2.$$

This implies that

$$\frac{b^2}{a^2}x^2 + y^2 = b^2.$$

Dividing both sides of this equation by b^2 we have

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Notice that here the x, y axes are not explicitly drawn but they are used implicitly.

3.4.4 Solution to higher degree equations in *The Geometry.*

The Third Book of *The Geometry* has the title *On the Construction of Solid and Supersolid Problems.* This chapter establishes our normal notation for higher degree equations. It also lays the foundation for polynomial algebra. Descartes' first order of business is to make clear that only positive roots of equations are true. We quote: *It often happens, however, that some of the roots are false or less than nothing.* He gives the example, first considering $(x-2)(x-3)(x-4) = x^3 - 9x^2 + 26x - 24$ with roots $2, 3, 4$. He then multiplies by $x + 5$ that has false root 5 (notice a false root in his sense is still described by a positive number and labeled as false). Thus he has

$$x^4 - 4x^3 - 19x^2 + 106x - 120$$

which has three true roots $2, 3, 4$ and one false root 5. He uses this as an example for his celebrated *Law of Signs.* Which says (the assertions refer to the signs occurring before the coefficients the coefficient of x^4 should be taken as $+1$).

An equation has as many true roots as it contains changes of sign from + to - or from - to +; and as many false roots as the number of times two + or two - are found in succession.

We note that 0 should be ignored. Thus the quartic example above the signs are $+, -, -, +, -$ thus it has 3 changes $+ \to - \to + \to -$ and two $-$ in succession so the theorem asserts that the number of true roots (i.e positive) is 3 and the number of false (negative) is 1. Thus the theorem is correct without any interpretation in this case. In general, there are some caveats. First we must count a root with multiplicity since $x^2 - 2x + 1 = 0$ has signs +,-,+. Hence 2 sign changes. But it only has one root 1. The method that he finds the example by successively multiplying seems to be all he feels is necessary for a proof of his assertion. In fact, proofs are noticeably absent from Descartes' book. However, there are several detailed derivations of formulas. In this case one can easily see that the above assertion is false as stated. If we consider

$$x^2 - 2x + 2 = 0$$

then the sequence of signs is +,-,+ so he predicts two true roots and no false ones. However there are no real roots of this equation by the quadratic formula. We must therefore interpret the result to be about equations with all roots true or false.

The next really substantive part of this chapter involves what happens when one starts with a polynomial in an unknown x and substitutes $x = y - a$ (in fact he uses the example of $a = 3$ in a variant of the polynomial he has been studying. He says that he has increased by 3 every true root and decreased by 3 every false root that is greater than 3. He then gives a full discussion of how to carry out the substitution on a specific quartic. He then considers the same calculation but this time diminishing by a that is substituting $x = y + a$ (again $a = 3$ and he looks at a specific quartic). These calculations take up a full half page of this extremely terse book. But then we come to the point he says:

> *...we can remove the second term of an equation by diminishing its true roots by the known quantity of the second term divided by the number of dimensions of the first term, if these terms have opposite signs; or if they have like signs by increasing the roots by the same quantity.*

In other words the reduction used in Cardano's *Ars Magna*. Descartes' formulas look just like ours but his text is still far from our approach of considering negative numbers as "true" objects. It is also important to note that it is here that he considers equations with indeterminate coefficients (we know them but they are arbitrary).

He later applies these considerations to the cubic and for all practical purposes gives a derivation of Cardano's formula for the solution of the cubic geometrically but writes the formula in exactly the same way that we do. We note that Descartes attributes it to Ferro (he in fact says that *"...the rule, attributed by Cardan to one Scipio Ferreus ..."*).

In the rest of the book he considers equations of higher degree mainly 6 and gives methods of solving specific equations. This small book lays the foundation of a synthesis of algebra (polynomials) and geometry. It lays out the power of a notational scheme that has lasted through our time.

3.4.5 Exercises.

1. Prove Descartes' rule of signs for polynomials of degree 1,2,3,4 with only real roots.

2. In the derivation of the equation for the ellipse it is necessary to have $a > c$. However, if we had $a < c$ then we would have had to write $-b^2 = a^2 - c^2$. Follow

the line of argument from there to see that

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

Can you make any sense out of this?

3.5 Higher order equations.

As we indicated Abel proved that, in general, an equation of degree 5 or higher cannot be solved using algebraic operations combined with extraction of roots (i.e. solvable by root extraction). Galois gave a method of determining which equations could be solved. Before we study what these two prodigies actually did in later chapters. In this chapter we will content ourselves with a better understanding of their accomplishment by improving our understanding of the concept of number. We will also resolve some ambiguities that arise in Cardano formula when we include (as we must) complex numbers. We will not attempt, as yet, to be completely rigorous (perhaps we never will) with this concept but will build on Descartes' ideas. We first introduce the concept of complex number more carefully than we did when we studied Bombelli's explanation of Cardano's strange example.

3.5.1 Complex numbers.

To Descartes, once a unit square is chosen the square of a number a, a^2 must be considered to be the area of the square of side a. Thus the square of a number can never be negative. However, in the Cardano formula we must include the possibility of taking the square root of a negative number. We note that if we wish to allow for this possibility we must only find a meaning for $\sqrt{-1}$ since if $a < 0$ then $a = -b$ with $b > 0$. So square root of a could be taken to be $\sqrt{-1} \times \sqrt{b}$. Thus if we wish to allow square roots of any number we need only make up a symbol for $\sqrt{-1}$. Engineers generally use j and mathematicians use i (for imaginary no doubt). Since there is no real number with our desired property we must "throw in" our new number i. Now we have a more complex type of number that looks like $c = a + bi$. We would like to maintain the rules of arithmetic so we are forced into

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

and

$$\begin{aligned} (a + bi)(c + di) &= (a + bi)c + (a + bi)di = \\ ac + bci + adi + bdi^2 &= (ac - bd) + (bc + ad)i. \end{aligned}$$

In other words with only our symbol i thrown in we can with apparent consistency define an addition and a multiplication. If we assume (as we must) that there is no relation of the form $a + bi = 0$ with a or b non-zero then we have a system that is consistent with arithmetic. We also note that if a or b is not 0 then

$$(a + bi)(a - bi) = a^2 - (bi)^2 = a^2 - (i)^2(b^2) = a^2 + b^2 > 0.$$

Thus if we set $\overline{a + bi} = a - bi$. Then if $c = a + bi$, $c\overline{c} = a^2 + b^2$ thus

$$\frac{\overline{c}}{c\overline{c}} = \frac{\overline{c}}{a^2 + b^2}.$$

So

$$\frac{\overline{c}}{a^2 + b^2} \times c = \frac{c\overline{c}}{a^2 + b^2} = 1$$

This tells us that if $c \neq 0$ then $\frac{1}{c}$ exists and is given by $\frac{\overline{c}}{a^2+b^2}$. We now have a number system that contains the square root of every real number. Let us call (as does everyone else) these numbers *complex numbers*. We assert that every complex number has a square root. In fact, the Fundamental Theorem of Algebra (mentioned earlier) asserts that every non-constant polynomial with complex coefficients has at least one root. The proofs of this theorem involve a deeper understanding of numbers than we have as yet and we will defer this to the next chapter where we will come to grips with the problem of rigorously explaining numbers. We will content ourselves, in this chapter, to showing that every complex number has n-th roots for all $n = 2, 3, ...$ For this we need trigonometry.

3.5.2 Exercises.

1. Show that $(1 + i)^2 = 2i$.

2. If a and $b \neq 0$ are given real numbers and if $c = \sqrt{\frac{a+\sqrt{a^2+b^2}}{2}}$ and $d = \frac{b}{2c}$ then show that $(c + di)^2 = a + ib$.

3. Use the formula in 2. to calculate $\sqrt{i}$.

3.5.3 Trigonometry.

We have seen in our discussion of Euclid's *Elements* that the Greeks were very interested in the properties of circles. They also had a notion of angle and studied methods of bisecting and trisecting angles. Our trigonometry is based on the understanding that angles can be represented by points on the unit circle. This is motivated by the following calculation. Consider $(x + iy)(u + iv) = (xu - yv) + (xv + yu)i = t + is$. We calculate

$$t^2 + s^2 = (xu - yv)^2 + (xv + yu)^2 = x^2u^2 - 2xyuv + y^2v^2 + x^2v^2 + 2xyuv + y^2u^2 =$$

$$x^2u^2 + y^2v^2 + x^2v^2 + y^2u^2 = (x^2 + y^2)(u^2 + v^2).$$

The conclusion we have been aiming at is that if $x^2 + y^2 = 1$ and $u^2 + v^2 = 1$ then the point that corresponds to $(x+iy)(u+iv)$ also has this property. If we *define* the unit circle to be the set of all complex numbers $z = x + iy$ such that $x^2 + y^2 = 1$. Then we conclude that the product of two elements of the unit circle is on the unit circle. We also note that if $y \geq 0$ and $x = iy$ is on the unit circle then $y = \sqrt{1 - x^2}$. If $y < 0$ then $y = -\sqrt{1 - x^2}$. Thus up to the sign we have parametrized the unit circle in terms of the value of the x coordinate and a sign. We are looking for a "better" parametrization in terms of a parameter θ. We wish to have $z(\theta) = x(\theta) + iy(\theta)$ with $z(\theta_1)z(\theta_2) = z(\theta_1 + \theta_2)$. If we multiply out we have

$$x(\theta_1)x(\theta_2) - y(\theta_1)y(\theta_2) = x(\theta_1 + \theta_2)$$

and

$$x(\theta_1)y(\theta_2) + x(\theta_2)y(\theta_1) = y(\theta_1 + \theta_2).$$

Also we have assumed that

$$x(\theta)^2 + y(\theta)^2 = 1.$$

There is an amazing fact that we will prove in our discussion of calculus. It says that if we have two functions of a real parameter satisfying the above three conditions and one more that asserts that if we make a small change in the value of θ then this induces a small change in the value of each of $x(\theta)$ and $y(\theta)$ then there is a fixed real number c such that $x(\theta) = \cos(c\theta)$ and $y(\theta) = \sin(c\theta)$. This is why the first two equations look so familiar. For the moment we will assume that we are all experts in trigonometry. We have therefore observed that the points of the unit circle can be described as $z(\theta) = \cos(\theta) + i\sin(\theta)$. We also know that there is a number π with the property that if we consider the values $z(\theta)$ for $0 \leq \theta < 2\pi$ then every point of the circle has been parametrized with a unique parameter. We also note that if have set up our parameter so that we traverse the circle counter-clockwise then we most have $z(0) = z(2\pi)$. Then using the property

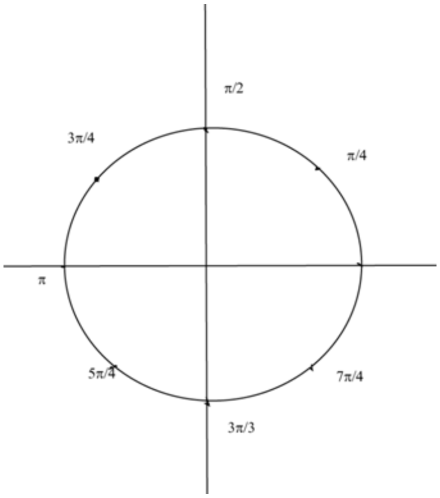
$$z(a + b) = z(a)z(b)$$

we must have $z(0)^2 = z(0)$. Now this implies $(z(0) - 1)z(0) = 0$. Since $z(0)$ is not zero (its on the unit circle). We must have $z(0) = 1$.

We can now make an observation due to Abraham De Moivre (1667-1754) (and perhaps to Jean d'Alembert (1717-1783)). If z is a complex number then we can write $z = rz(\theta)$ with $r = \sqrt{x^2 + y^2}$ and $z(\theta) = z/r$. We presume that we can take arbitrary roots of non-negative real numbers. So if we want an n-th root of z we can take $r^{\frac{1}{n}}z(\frac{\theta}{n})$. This says that a complex number has at least one n-th root for each n. We have seen that a positive real number has two square roots $\pm\sqrt{x}$ the square-root symbol always stands for the non-negative square root. The point here is the square roots of 1 are ± 1. The same sort of thing happens in general. If $a^n = b^n = z$ then $(a/b)^n = 1$. Thus the ambiguity in taking roots is contained in the n-th roots of unity. Here is the observation:

$$z(\frac{2\pi k}{n})^n = z(2\pi k) = z(2\pi)^k = z(0)^k = 1^k = 1.$$

This is explained in terms of the following picture (here we have plotted 8 equally spaced points on the circle $\frac{2\pi k}{8}$ with $k = 0, 1, 2, 3, 4, 5, 6, 7$)



We can see that we are just putting n (in this place 4) equally spaced points on the circle with the first one 1. Multiplication by the second one clockwise cycles the points clockwise around the circle.

Interpretation of Cardano's formula. We now see that there is a real problem with Cardano's formula (and Ferrari's for that matter). In Cardano there are two cube roots and two square roots thus there is a possible thirtysix-fold ambiguity in the formula. The only way to make it a formula again is to give a rule for how to choose the roots. Let's look at the equation

$$x^3 = ax + b$$

again. The formula says

$$\sqrt[3]{\frac{b}{2} + \sqrt{\left(\frac{b}{2}\right)^2 - \left(\frac{a}{2}\right)^3}} + \sqrt[3]{\frac{b}{2} - \sqrt{\left(\frac{b}{2}\right)^2 - \left(\frac{a}{2}\right)^3}}$$

is a root. We note that the formula involves both square roots of $\left(\frac{b}{2}\right)^2 - \left(\frac{a}{2}\right)^3$ symmetrically. So the ambiguity involves only how we choose cube roots. We will therefore use the same square root, v,of $\left(\frac{b}{2}\right)^2 - \left(\frac{a}{3}\right)^3$ in both parts of the formula. We write $u = \frac{b}{2}$. Then we must choose a cube root α of $u+v$ and a cube root β of $u-v$ so that $x=\alpha+\beta$ is a solution to the equation. Let us calculate. With α,β arbitrary choices of cube roots. Then $x^3=\alpha^3+3\alpha^2\beta+3\alpha\beta^2+\beta^3$. We therefore have

$$x^3=u+v+3\alpha^2\beta+3\alpha\beta^2+u-v=b+3\alpha^2\beta+3\alpha\beta^2.$$

Now $3\alpha^2\beta+3\alpha\beta^2=3\alpha\beta(\alpha+\beta)$. We note that $\alpha^3\beta^3=(u+v)(u-v)=u^2-v^2=\left(\frac{b}{2}\right)^2-\left(\left(\frac{b}{2}\right)^2-\left(\frac{a}{3}\right)^3\right)=\left(\frac{a}{3}\right)^3$. Thus to help resolve the ambiguity of cube roots we choose α and β so that $\alpha\beta=\frac{a}{3}$. (Notice that we can do this by multiplying one of α or β by a complex number whose cube is 1.) We now note that $\alpha+\beta=x$ (by our definition) and $3\alpha\beta=3\frac{a}{3}=a$. Thus $3\alpha^2\beta+3\alpha\beta^2=ax$. With these choices and $x=\alpha+\beta$ then the equation $x^3=ax+b$ is satisfied. We also note that since the choice of α forces that of β and vice-versa there is now only a threefold ambiguity. Which is what we would have if $a=0$. The "general" polynomial of degree three has 3 roots.

3.5.4 Exercises.

1. What are the 4 eighth roots of 1?

2. Find a fourth root of -1 and show that its powers give up to reader the 8 equally spaced points around the circle in the picture above.

3. Resolve the ambiguity in Cardano's formula for a solution of $x^3+ax=b$.

4. Write out the three roots of the equation $x^3+2x=4$.

3.5.5 Polynomials of degree 5 or higher.

We will begin this section with a special case of Gauss's fundamental theorem of algebra. A fuller explanation of the argument in the next subsection will be given in the next chapter. Also the following theorem will be an ingredient in our development of the full theorem. We will see that the result is based on a deeper understanding of the concept of a real number. We will be studying the full fundamental theorem of algebra in the next chapter. Here we will give an argument for polynomials with real coefficients of odd degree that uses methods of analysis (the subject of the next chapter). The proof involves a deep property of real numbers which we will assume. The reader who has not had any introduction to the manipulation of inequalities might find that the proof below is gibberish. Try reading it anyway. The mysteries will be expanded on in the next chapter.

Polynomials of odd degree. The purpose of this subsection is to discuss the following

Let $f(x) = a_0 + a_1x + ... + a_nx^n$ be a polynomial with $a_n \neq 0$, $a_0, ..., a_n$ real numbers and n odd. Then there exists a real number c such that $f(c) = 0$.

Notice this assertion is not about arbitrary polynomials but only ones of odd degree and having real coefficients. In particular, if the coefficients are rational then there is a real root. As we observed above this result is a consequence of a deep property of real numbers which will be delved into more deeply in the next chapter. We first note that we may assume that $a_n = 1$ since we can divide through by a_n. Thus we are looking at

$$f(x) = x^n + (a_0 + a_1x + ... + a_{n-1}x^{n-1}).$$

We assume that $C > |a_i|$ for all $i = 0, ..., n-1$. Then $|a_0+a_1x+...+a_{n-1}x^{n-1}| \leq |a_0| + |a_1||x| + ... + |a_{n-1}||x|^{n-1}$. Then

$$|a_0 + a_1x + ... + a_{n-1}x^{n-1}| \leq C + C|x| + ... + C|x|^{n-1}.$$

Thus if $|x| > 1$ then we have

$$|a_0 + a_1x + ... + a_{n-1}x^{n-1}| \leq nC|x|^{n-1}.$$

We now note that if x is real then $f(x) \leq x^n + nC|x|^{n-1}$. Now suppose that $x < 0$ and $|x| > 2nC$. Then since n is negative $x^n = -|x||x|^{n-1}$. Thus

$$f(x) \leq -|x||x|^{n-1} + nC|x|^{n-1} < -2nC|x|^{n-1} + nC|x|^{n-1} = -nC|x|^{n-1} < 0.$$

We conclude that if $x < 0$ and $|x| > 2nC$ then $f(x) < 0$. We note that if a and b are real numbers then $a + b \geq a - |b|$. Indeed if $b \leq 0$ then this is an equality and if $b > 0$ then $b > -|b|$. Thus if $x > 2nC$ and $x > 1$ then

$$\begin{aligned} f(x) &\geq x^n - |a_0 + a_1x + ... + a_{n-1}x^{n-1}| \geq x^n - nC|x|^{n-1} \\ &= |x||x|^{n-1} - nC|x|^{n-1} \geq 2nC|x|^{n-1} - nC|x|^{n-1} = nC|x|^{n-1} > 0. \end{aligned}$$

We have thus shown that if $x < -nC$ and $x < -1$then $f(x) < 0$ and if $x > 2nC$ and $x > 1$ then $f(x) > 0$. Fix $u < -1$ and $u < -nC$. Fix $v > 1$ and $v > 2nC$. Then $f(u) < 0$ and $f(v) > 0$. The deep property that we need is that if g is a polynomial and if we have two real numbers $a < b$ then for every real number, c, between $g(a)$ and $g(b)$ there exits a real number, y, with $a \leq y \leq b$ such then $g(y) = c$.

We apply this to f. Then since $f(u) < 0$ and $f(v) > 0$ the number 0 is between $f(u)$ and $f(v)$ and thus there exists y with $u \leq y \leq v$ and $f(y) = 0$.

The property we have used is called the *intermediate value property* and it applies in much greater generality than polynomials (as we shall see in the next chapter).

Why haven't we found a paradox? In the previous subsection we have observed that if we have a polynomial of odd degree with real coefficients then it has a real root. The existence of the root is demonstrated using a deep property of the real numbers not by giving a formula. This leads to the question: If we have a polynomial of degree 5 with rational coefficients then what is the nature of the real numbers that are its real roots? The point is that for degrees $2, 3$ and 4 the roots were found it the collection of complex numbers that can be found by the following operations on rational numbers:

1. Arithmetic (addition, subtraction, multiplication and division).

2. Extraction of roots ($\sqrt[2]{x}, \sqrt[3]{x}, ...$) which we now understand how to do using trigonometry.

For example in Cardano's formula we must extract a square root of an expression involving the coefficients of the polynomial, do arithmetic with that not necessarily rational number combined with a further coefficient and then extract cube roots and then subtract these numbers.

The amazing outcome of the work of Ruffini, Abel and Galois is that these operations are not enough to find all roots of polynomials with rational coefficients of degree at least 5. This goes far beyond showing that we cannot find an explicit formula using only operations of type 1. or 2. It shows that roots of polynomials with rational coefficients form an algebraic object that is much more subtle than was imagined. In the next chapters we will endeavor to explain the analysis that is involved in Gauss's fundamental theorem of algebra. This analysis comes from the foundations of the differential and integral calculus which had been developed for totally different purposes (the determination of velocities, accelerations and tangents). Also the study of roots of equations led to what is now called abstract algebra. The abstraction of addition, multiplication and division. The whole is a startling edifice that will be one of the main subjects of the remaining chapters.

We will content ourself with a discussion of why the Abel, Ruffini, Galois theory was a surprise to the mathematicians of the eighteenth and nineteenth centuries (as the theory developed). This involves the question of why mathematicians with only the knowledge that there is a formula for a solution of an equation of degree two with (say) rational coefficients seemed to expect that there would be analogous formulas in higher degrees? Since it apparently took mankind about 4000 years from the realization that there was a quadratic formula to the time of Cardano and Ferrari when formulae for degrees 3 and 4 were discovered, why were mathematicians not looking for reasons why this couldn't be done? The expectation that it could, in fact, be done was correct for degrees 3 and 4 but definitely incorrect for higher degree. This time it took approximately 200 years to come to the realization that one could not do for degree 5 what was done for 2,3 and 4. This is not unlike the prevailing feeling of mathematicians before the nineteenth century that the parallel postulate was a consequence of the other axioms of Euclid's geometry. The point is that the mathematicians had decided that they believed the validity of a statement that they could not prove. Since there was no justification for this belief one should perhaps call it a prejudice. In the latter part of the twentieth century and the beginning of this (the twenty first) century there has been a new debate on the question of truth without proof Brilliant expositors of mathematics justify their views of computability and artificial intelligence with just such an idea. Indeed, the argument is that if a human being can discern the truth of an assertion without a proof then he can find true statements that could never be found by a computer. Thus human beings must be more than biological computers. We will not enter this fray which is poised at a higher level than we have scaled as yet. Rather, the history of the search for formulae for the solution of polynomial equations and the prejudice that this could be done is perhaps related to a problem with the flexibility of the human mind. That so many believe that assertions must be true even though we can't prove them may be related to the prejudices that were at the root core of the horrible events of the twentieth century.

3.5.6 Exercises.

1. Does the intermediate value property: *If a polynomial, $f(x)$ with real coefficients, takes two values $f(a) > 0$ and $f(b) < 0$ then there exists a real number, c, between a and b such that* $f(c) = 0$. Seem obvious? Is it it true if we replace the word real by rational?

2. Let $f(x) = x^5 + 2x^2 + x + 1$ show that there is a real root between -1.23 and -1.22 by calculating the two values and seeing that one is negative and the other is positive. If you have access to a computer algebra package you could use it to check that the intermediate root does indeed exist.

4 The dawning of the age of analysis.

Archimedes (287-212 BC) proved that the area of a circle is equal to the area, A, of a right triangle with one side equal in length to the radius and the hypotenuse equal to the circumference (see in discussion in Chapter 2). His method was to observe that the area of the triangle in question is either equal to, strictly less than or strictly greater then the area of the circle. He then inscribes regular polygons with the number of sides increasing indefinitely and shows that they are eventually bigger in area than any number strictly less than A he then circumscribes regular polygons of increasingly many sides and shows that eventually they have area less than any number strictly larger than A. He then concludes that the only possibility left is that A is the area. This type of argument replacing direct calculation with upper and lower bounds is the method of modern analysis. The main impetus for the development of a rigorous branch of mathematics which we now call analysis was the need for a consistent underpinning for (what we now call) Calculus (Isaac Newton(1642-1727), Gottfried Leibniz(1646-1716). The term calculus is a generic term that roughly means "a method of calculation". It was a revolutionary idea that led to simple methods of calculating areas and tangents in geometry and velocities, accelerations and trajectories in mechanics. In particular, the clever method of Archimedes becomes unnecessary within the framework of Calculus. Unfortunately, the early methods were completely formal implicitly assuming that one can deal with quantities that were so small that their squares could be treated as 0 (fluents in the terminology of Newton, infinitesimals to others). Although there was no rigorous notion of infinitesimal in the seventeenth and eighteenth century the idea led to such amazing simplifictions of difficult problems that the theory led to a revolution in mathematics. As we shall see in later chapters a more rigorous approach to solving the same problems was developed in the nineteenth century and was based on the ideas of "modern analysis". We wiill see in this chapter that Fermat (1601-1665) had developed methods consistant with modern analysis to compute certain important areas and tangents. But he had no general calculus based on modern analysis. In the twentieth century a more rigorous version of the infinitesimal calculus was developed by Abraham.Robinson(1918-1974) based on a deep understanding of logic which made the formal methods of the seventeenth and eighteenth centuries more acceptable in the twentieth centuries.

All attempts at understanding a firm basis of calculus are in the end based on attempts to understand the real number system. This was part of our goal in the previous chapter. There, we showed how Euclid and Descartes had developed numbers out of geometry. Descartes went much further and showed that the algebraic manipulation of numbers could replace the clever methods of geometry. However, Descartes' numbers did not have any existance beyond geometry. We also saw that the basic question of whether ther exist roots of polynomial equations and whether or not we can calculate them also devolves on the question: "What is a real number" and thereby "what is a complex number". We will be studying these points in this chapter. The modern formulation of real and complex numbers will have to wait for the next chapter.

4.1 Early aspects of analysis.

4.1.1 Zeno's paradox.

We will begin this chapter with a standard puzzle usually attributed to Zeno (490-425 B.C.). Suppose there were a tortoise and a hare (sometimes it is Achilles) such that the hare moves twice as fast as the tortoise. To simplify things we assume that the tortoise can move 1 unit in a second and the hare can move 2 units in a second. Suppose that the tortoise starts moving first along a straight line and travels a distance d before the hare begins moving along the same line. We then have the following situation in the first $\frac{d}{2}$ seconds: the hare is d units from the starting point and the tortoise is $\frac{3d}{2}$. In the next $\frac{d}{4}$ seconds the hare has moved to $\frac{3d}{2}$ units and the tortoise is at $\frac{7d}{4}$, that is the tortoise is still ahead by $\frac{d}{4}$ units. After the next $\frac{d}{8}$ seconds the tortoise will be ahead by $\frac{d}{8}$ units, etc. Thus the hare will never catch the tortoise! We know that there is something wrong here since it is obvious that the hare will eventually pass the tortoise.

Aristotle (384-322 B.C.) used this "paradox" as evidence for the premise that infinity is meaningless. This is, certainly a practical point of view. We cannot do an infinite number of operationseach of which take at least a fixed amount of time to accomplish. But this is not what is happening in our discussion of the tortoise and the hare. If we redid the steps and did the measurement in fixed units of time, say one second. Then after n seconds the tortoise would be at $d+n$ units from the start and the hare would be at $2n$ units. Thus after (say) $d+\frac{1}{2}$ seconds the tortoise would be at $2d+\frac{1}{2}$ units from the start and the hare would be $2d+1$ units. That is they will pass each other before $d+\frac{1}{2}$ seconds elapse. Let's look at our original analysis let us make the problem more concrete by taking d ot be 100. Then after 10 steps the tortoise is $\frac{100}{1024}$ units ahead of the hare. After 20 interations of this procedure it is $\frac{100}{1048576}$ ahead. If say the units were meters then this is less than .0001 meters and the amount of time to travel that far for the tortoise would be that many seconds. This is absurd. There is no way we can measure that small an interval in time (let alone what we would have a few iterations further along). The time intervals are becoming so small as to be meaningless. However this is not a solution to the "paradox". For example, it is possible for the toroise to move as far as he wishes even if he moves in certain incriments of time that become arbitrarily small. Here we look at just the tortoise and look at where he is from the start after 1 seccond, a 1/2 half second later, a 1/3 second later,... Then after 2 such time intervals he would have gone $\frac{3}{2}=1.5$ units, after 4 he would have gone $\frac{25}{12}\sim 2.08$, after 8 it would be $\frac{761}{280}\sim 2.72$, after 16 it would be $\frac{2436559}{720720}\sim 3.3$,after 200 it would be about 5.88. after 10000 it would have gone 9.79 units. We will show that the numbers defined in this way increase without bound (see the section immediately below on the harmonic series). Thus just saying that the time incriments are becoming too small to measure does not resolve the puzzle. Many look upon this puzzle as indicating a need for a better understanding of infinity. We will take a different approach and explain how the techniques of modern analysis explain that the "puzzle" is merely a missunderstanding of the finite.

4.1.2 The harmonic series.

In this section we will use the method of Nicole d'Oresme (1323-1382) to show that the numbers $1+\frac{1}{2}+\frac{1}{3}+...+\frac{1}{n}$ increase without bound with n. The idea of Oresme can be seen as follows:

$$1+\frac{1}{2}=1+\frac{1}{2},$$

$$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4} > 1+\frac{1}{2}+\frac{1}{4}+\frac{1}{4} = 1+\frac{1}{2}+\frac{1}{2}$$

(here we have observed that $\frac{1}{3} > \frac{1}{4}$)

$$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8} =$$

$$1+\frac{1}{2}+(\frac{1}{3}+\frac{1}{4})+(\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}) >$$

$$1+\frac{1}{2}+\frac{1}{2}+(\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}) = 1+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}.$$

The pattern is now clear if we add up

$$\frac{1}{2^n+1}+\frac{1}{2^n+2}+...+\frac{1}{2^{n+1}}$$

We have 2^n terms that are all at least as big as $\frac{1}{2^{n+1}}$. Thus they add up to a number that is at least $2^n\left(\frac{1}{2^{n+1}}\right) = \frac{1}{2}$. The conclusion is that the sum

$$1+\frac{1}{2}+\frac{1}{3}+...+\frac{1}{2^n-1}+\frac{1}{2^n} \geq 1+\frac{n}{2}.$$

If we define the integral logarithm in base 2 by $I\log_2(N) = n$ if $2^{n-1} < N \leq 2^n$. Then we have

$$1+\frac{1}{2}+\frac{1}{3}+...+\frac{1}{n} \geq 1+I\log_2(n).$$

This beautiful argument actually gives a very good idea of how this series of numbers grows with n. One can show that there exists a constant (Euler's constant) that is usually denoted γ and another constant we will call μ which (we shall see just the natural logarithm of 2) such that if we substitute increasingly larger values of n in the expression

$$1+\frac{1}{2}+\frac{1}{3}+...+\frac{1}{2^n}-n\mu-\gamma$$

it becomes smaller than any preassigned (small number). We will discuss this in more detaiol when we talk about logarithms. This constant γ occurs in many contexts in mathematics and has been calculated to high precision. However, it is not known if it is a rational number.

Exercises.

1. Suppose you have blocks each 1 unit thick 4 units wide and 12 units long (the unit could be inches or cent1meters the dimensions are not terribly important) made of a uniform material. Suppose you were to pile the bocks one on top of each other so that the second overhangs the first the third overhangs the second, etc. How big an overhang could we achieve?

2. Use a computer algebra package or calculater with high precision and natural logarithms (ln)to calculate

$$1+\frac{1}{2}+\frac{1}{3}+...+\frac{1}{n}-\ln(n)$$

for large n. What is the value of γ that your calculation predicts.

4.1.3 Another look at the methods of Archimedes.

As we saw Archimedes developed a method of proving formulas for areas of geometric figures. His method was to have a target value, A, for the area in mind and to show that for any $B > A$ there exists a geometric figure strictly containing the one in question with area that we know how to compute and which is less than B. He then showed that if $C < A$ then there was a figure stricly inside the one in question with area bigger than C. He than concludes that the area is bigger than any number strictly bigger than A and is less than any number strictly bigger than A. He then asserts that this implies that the area must be A. The argument is ingenious but once understood seems self evident. However, there are several (reasonable) assumptions that have been made and there is one problem with the "method". We will first look at the assumptions. The first is about numbers (which eventually leads to the notion of a Dedekind cut) the then there are two about areas. The assumption about numbers is:

1. If A and D are numbers then $A = D$ if the following two conditions are satisfied
a) Every B satisfying $B > A$ also satsifies $B > D$.
b) Every C satsifying $C < A$ also satsifies $C < D$.

The first assumption about areas is:
2. If F and G are subsets of the plane with areas A and B and if every point in F is also in G then $A \leq B$.

It is hard to disagree with these two assumptions. The next is not clearly an assumption at all will eventually become part of the definition of a set with area.

3. Let F be a subset of the plane. Suppose that A is a number such that whenever G is a set that has an area B that contains F then $B \geq A$ and whenever L is a set that is contained in F and has an area C then $C \leq A$. Then F has area A.

The first "assumption" is part of the order properties of the real numbers. The second is a property that must be satisfied if we are to have a reasonable notion of area. The third has to do with the fact that in the contexts that are least "weird" to mathematicians not every set can be allowed to have an area with condition 2. (and a few more equally "obvious" conditions) satisfied.

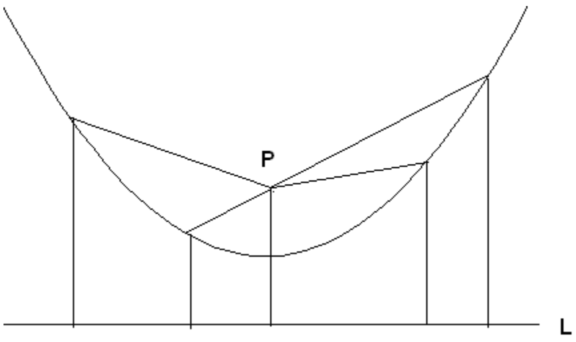
We also indicated that there was a problem with Archimedes method. The problem is that it is a method that proves that a value asserted for an area is the correct one. It gives no method of finding what the value should be. We know that Archimedes was aware that the area of the circle of radius r is πr^2. He also clearly knew that the circumference is $2\pi r$. The right triangle with sides of length r and and the other of length the circumference has area

$$\frac{1}{2}r \cdot (2\pi r) = \pi r^2.$$

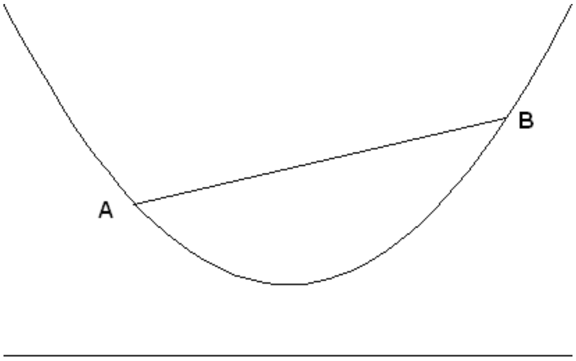
Thus he has proved the formula that we believe. (Actually what he has done is reduced the problem of calculating the area to the problem of calculating the circumference or vice-versa.) A general method for calculation was one of the main aims in the development of the infinitesimal and integral calculus.

Archimedes was the first to calculate the area of a segement of a parabola. We will not go into his derivation but say that his basic axioms of area were also used and the area was given in terms of the area of a triangle that one can only feel was an outgrowth of an amazing insight. After we discuss calculus we will explain Archimedes remarkable formula.

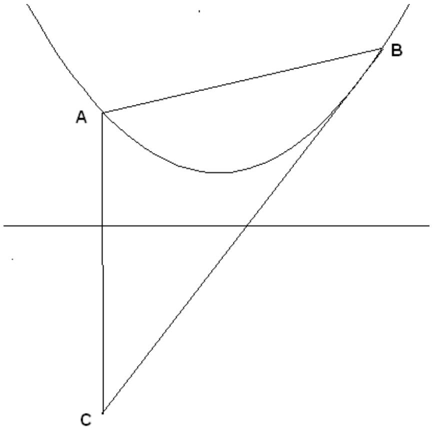
Recall that a parabola is a conic section that is determined by a point, P, and a line, L. The curve is the locus of points whose distance to P is equal to its distance to the line L. For instance



The point P is called the *focus* of the parabola and the line perpendicular to L through P is called the *axis* of the parabola. The Greeks understood that for a conic section a line could intersect it in two, one and no points. A line that intersected the curve at one point would be called the tangent line to that point (it was known to be unique). Archimedes in ??? set about to calculate the area of what he called a section of the parabola, that is,the set is cut out by a line intersecting with the parabola at two points A and B.



He assumes that the point A and B are on different sides of the axis and A is closer. He then considers the triangle ABC formed by the tangent line through B the segment AB and the line parallel to the axis of the parabola through A.



The theorem of Archimedes is:

The area of the parabolic segment is one third the area of the triangle ABC.

This theorem is one of the high points of Greek geometry. Archimedes approach (as we have pointed out) involved a guess of the area and then trhough brilliant upper and lower estimates proving that his asserted area is correct. In fact, he had a method of deciding what the appropriate value should be that involved what he called *The Method*. This method was based on what is now called *statics* in physics involving the theory of levers and pulleys. So in addition to being one of the greatest mathematicians who ever lived he was also a great physicist. The story of how *The Method* was rediscovered after seeming to be lost for over a thousand years is also very interesting. We refer to more standard texts in the history of mathematics for this story (e.g. C. Boyer et. al. *A History of Mathematics*).

Exercises.

1. Derive an equation for the parabola using the plane coordinates (x, y) if we take the line L to be given by $y = -\frac{1}{4}$ and the focus to be the point $(0, \frac{1}{4})$.

2. For the parabola in problem 1. show that if the line AB is parallel to the axis then the endpoints are given with A having x-coordinate $-a$ and B having x-coordiane a and the area of the indicated triangle is $4a^3$ so the area of the sector is $\frac{4a^3}{3}$.

3. Show that the theorem of Archimedes shows that the area of the parabolic segment AB depends only on the sum of the distances of A and B to the axis.

4.2 Precursors to calculus.

As mentioned above, Nicole d'Oresme had developed methods for studying the growth of infinite sequences of numbers. He also understood fractional powers of positive numbers and most astonishingly used graphical methods to plot data (using a horizontal axis for the independent variable and the verticle axis for dependent variable. However, very little progress was made in the years between Archimedes and Oresme in the calculation of areas bounded by curves. One major drawback was that the mathematical notation was still quite cumbersome and the methods of Oresme to visualize were not widely used.

In the last chapter we mentioned the work of Viéte which explained how to deal with unknown quantities and thereby led to the concept of function. However, his work did not separate the notion of number from geometry. Thus, positive numbers were lengths of intervals, products of positive numbers were areas of rectangles, triple poducts were volumes, etc. He also used cumbersome notation for powers writing something like $x$cube for what we would write as x^3. Thus he would have x $x$square = $x$cube. This did not afford a useful formalism for doing algebraic manipulation of polynomials. In our notation, a polynomal:

$$x^3 + 2x^2 + 5x + 1$$

would be understood to represent a volume so the 2 would would be in units of length, the 5 in units of area and the 1 would be a volume. You could then visualize a cube of side x a rectangular box with base of side x and height 2, a rectangular box with base of area 5 and height x and a three dimensional figure of volume 1 all attached to each other in some way.

This all changed with the French mathematicians of the first half of the seventeenth century. We have already written about Descartes' explanation of how to interpret products of positive numbers as intervals and thereby freed polynomials of units. Of the great mathematicians of this time the one who arguably came the closest to calculus was Fermat. He, in fact, was more in the tradition of Archimedes than the later formal methodology of calculus. That is, more in line with the modern notion of analysis.

4.2.1 The Pascal triangle and the Leibniz harmonic triangle.

We recall that $(x+y)^2 = x^2+2xy+y^2$. Multiplying out we see that $(x+y)^3 = x^3+3x^2y+3xy^2+y^3$. We can continue to multiply indefinitely and we see that $(x+y)^n$ has $n+1$ terms the ith a multiple of $x^{n-i}y^i$. If we lay out the coefficients we have for $n = 0, 1, 2, 3, 4, 5$.

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}$$

Blaise Pascal (1623-1662) observed the pattern that one had a triangle with the two legs all ones and the interior values gotten by adding together the two adjacent values one row up for the interior points. Thus for the fifth power we would get $1, 5, 10, 10, 5, 1$ and, say, $10 = 4+6$. The standard method of writing these coefficients is $\binom{n}{i}$ so $\binom{5}{2} = 10$. With the conventions that $\binom{n}{i} = 0$ if $i < 0$ or $i > n$. It is convenient to write $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ (these account for the outer legs of the triangle). We have

$$(x+y)^n = x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + ... + \binom{n}{n-1}xy^{n-1} + y^n.$$

If we multiply this identity by $(x+y)$ then we have

$$(x+y)x^n+\binom{n}{1}(x+y)x^{n-1}y+\binom{n}{2}(x+y)x^{n-2}y^2+...+\binom{n}{n-1}(x+y)xy^{n-1}+y^n.$$

Now $\binom{n}{i}(x+y)x^{n-i}y^i = \binom{n}{i}x^{n+1-i}y^i + \binom{n}{i}x^{n+1-i-1}y^{i+1}$. This says that in the product the coefficient of $x^{n+1-i}y^i$ is

$$\binom{n}{i-1}. + \binom{n}{i}$$

Which is the pattern Pascal observed. We will call this the *generating identity* for the binomial coefficients. We note that if $0 \leq i \leq n$ then $\binom{n}{i} \neq 0$.

Some years after Pascal's discovery Christiaan Huygens (1629-1695) asked Leibniz to sum the series:

$$\frac{2}{1(1+1)} + \frac{2}{2(2+1)} + ... + \frac{2}{n(n+1)} + ...$$

that is the sum of the reciprocals $\frac{1}{\binom{n+1}{2}}$, $n = 1, 2, 3, ...$ The rigorous theory of summing such series had not as yet been developed. However, Leibniz came up with as solution that incontrovertibly summed the series to 2. Here is what he did. He observed that

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}.$$

This says that if we sum the first, say, 5, terms we have

$$2\left(1-\frac{1}{2}\right)+2\left(\frac{1}{2}-\frac{1}{3}\right)+2\left(\frac{1}{3}-\frac{1}{4}\right)+2\left(\frac{1}{4}-\frac{1}{5}\right)+2\left(\frac{1}{5}-\frac{1}{6}\right)=2-\frac{2}{6}.$$

If we sum the first n terms we get $2-\frac{2}{n+1}$. Thus if we sum a million terms the sum is 2 to 5 significant figures. The more terms we add the closer the value is to 2. This is essentially the modern version of sumation of infinite series. We should, however, point out that one must be careful about formal manipulation of infinite series. For example, suppose we want to sum

$$1-1+1-1+1+...$$

If we sum the first $2n$ terms we get $(1-1)+(1-1)+...+(1-1)=0$. If we sum the first $2n+1$ terms we get $(1-1)+(1-1)+...+(1-1)+1=1$. Leibniz felt that a reasonable value for the sum of this series should be $\frac{1}{2}$.

Returning to the reciprocals of the binomial coefficients $\binom{n+1}{2}$ Leibniz made a beautiful discovery that allowed him to compute many more infinite series using exactly the same trick.. He first observed Pascal's triangle could be written somewhat differently as

$$\begin{array}{cccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & \cdots \\
1 & 3 & 6 & 10 & 15 & 21 & 28 & \cdots \\
1 & 4 & 10 & 20 & 35 & 56 & 84 & \cdots \\
1 & 5 & 15 & 35 & 70 & 126 & 210 & \cdots \\
1 & 6 & 21 & 56 & 126 & 252 & 462 & \cdots \\
1 & 7 & 28 & 84 & 210 & 462 & 964 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

One observes that except for the first row and column if we look at an entry in this double array then it is the difference between the entry directly below and the entry one below and one to the left. Thus 6 in the third row has 10 directly below and four one down and one to the left We have $6=10-4$. To see this property is true we note that the entries in the first row are $\binom{0}{0},\binom{1}{0},\binom{2}{0},\binom{3}{0}$.... Those in the second row are $\binom{1}{1},\binom{2}{1},\binom{3}{1},\binom{4}{1}$... in the third $\binom{2}{2},\binom{3}{2},\binom{4}{2},\binom{5}{2}$,..., etc. Thus the entry in the i,j position is $\binom{j+i-2}{i-1}$. In other words the element in the $3,5$ positions is $\binom{8-2}{2}=15$ that in the $5,4$ position is $\binom{7}{3}=35$. The assertion above is just $\binom{j+i-2}{i-1}=\binom{j+i-1}{i}-\binom{j+i-2}{i}$ moving the negative term to the right hand side we see that this is the generating identity for the binomial coefficients. Now if we use the method of Leibniz above we find that if we consider an entry not on the first row or column and add up all the entries in the column directly to its left that are either on the same row or a higher row then we get the original entry . For example:

$$462=210+126+70+35+15+5+1.$$

The harmonic triangle of Leibniz is given by

$$\begin{array}{cccccccc}
1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \cdots \\
\frac{1}{2} & \frac{1}{6} & \frac{1}{12} & \frac{1}{20} & \frac{1}{30} & \frac{1}{42} & \frac{1}{56} & \cdots \\
\frac{1}{3} & \frac{1}{12} & \frac{1}{30} & \frac{1}{60} & \frac{1}{105} & \frac{1}{168} & \frac{1}{252} & \cdots \\
\frac{1}{4} & \frac{1}{20} & \frac{1}{60} & \frac{1}{140} & \frac{1}{280} & \frac{1}{504} & \frac{1}{840} & \cdots \\
\frac{1}{5} & \frac{1}{30} & \frac{1}{105} & \frac{1}{280} & \frac{1}{630} & \frac{1}{1260} & \frac{1}{2310} & \cdots \\
\frac{1}{6} & \frac{1}{42} & \frac{1}{168} & \frac{1}{504} & \frac{1}{1260} & \frac{1}{2772} & \frac{1}{5544} & \cdots \\
\frac{1}{7} & \frac{1}{56} & \frac{1}{252} & \frac{1}{840} & \frac{1}{2310} & \frac{1}{5544} & \frac{1}{12012} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}.$$

Here the first row consists of the numbers $1, \frac{1}{2}, \frac{1}{3,}, \frac{1}{4}, ...$ that is the terms in the harmonic series. The second row consists of the numbers $\frac{1}{n(n+1)}$ that is $\frac{1}{2}, \frac{1}{6}, \frac{1}{12}, \frac{1}{20}, ...$ The third row consists of the numbers $\frac{1}{3\binom{n+2}{3}}$ that is $\frac{1}{3}, \frac{1}{12}, \frac{1}{30}, \frac{1}{60}, ...$ The k-th row has entries $\frac{1}{k\binom{n+k-1}{k}}$ these entries can be read off from Pascal's triangle. Leibniz' observation was that the sum of the entries in the k-th row from the n-th poistion on is given by the number in the $k-1$-st row in the n-th position. Thus the sum

$$\frac{1}{60} + \frac{1}{105} + \frac{1}{168} + \frac{1}{252} + ... = \frac{1}{20}.$$

That is we are summing the entries in the third row starting with the fourth entry the sum is the fourth entry in the second row. In particular the answer to Huygen's question is twice the sum of the entries in the second row which is twice the first entry in the first row which is 2.

Although the Leibniz method was ingeneous it was a severely limited ,method of summing infinite series. A series that looked very similar to the series

$$1 + \frac{1}{3} + ... + \frac{2}{n(n+1)} + ...$$

is

$$1 + \frac{1}{4} + ... + \frac{1}{n^2} + ...$$

This series baffled Leibniz who was asked to sum it in 1673 by Henry Oldenburg (1615-1677) and, in fact, all mathematicians until Euler determined its sum in about 1736. We will come back to this series later in this chapter.

Exercises.

1. Use the harmonic triangle to sum the series

$$\frac{1}{6} + \frac{1}{24} + ... + \frac{1}{(n+2)(n+1)n} + ...$$

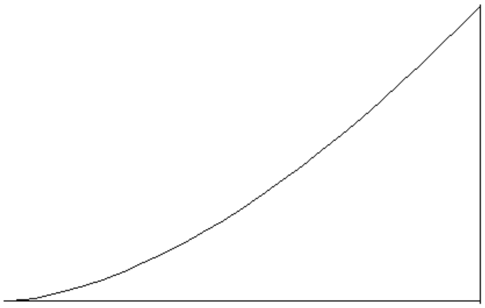
2. Use the method that Leibniz used in answering Huygens to show that the entire harmonic triangle works as advertised. Hint: we need to show that

$$\frac{1}{k\binom{n+k-1}{k}} - \frac{1}{k\binom{n+k}{k}} = \frac{1}{(k+1)\binom{n+k}{k+1}}.$$

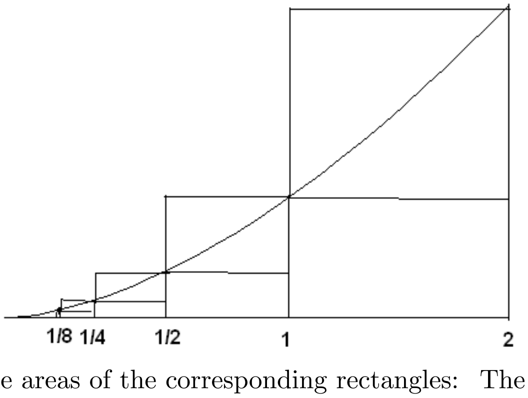
3. You may wonder why we called Leibniz' array a triangle. If you rotate the rectangular version of Pascal's triangle 45 degrees to the right (clockwise) then it it is a triangle. Do the same with the harmonic version. Explain the "geometyr" of summing series in terms of the version that is given as a triangle.

4.2.2 Fermat's calculation of areas.

Fermat considered the problem of calculating the area of a figure bounded by a line L_1, a line L_2 perpenducular to L_1 and a curve which we would write as $y = x^{\frac{p}{q}}$ with $p, q > 0$ relatively prime integers. He also allowed p to be negative but his method failed for $\frac{p}{q} = -1$. We will discuss this case later, although it was done chronologically earlier. In fact, the case of $q = 1$ had been handled by several authors who came before Fermat. Here is a picture of Fermat's area corresponding to the curve $y = x^{\frac{5}{3}}$ with the base of length 2.



Let the length be denoted m. His idea was as follows consider a number $0 < E < 1$ then one has the points $E^k m$ for $k = 0, 1, 2, ...$ Which start with m and decrease to indefinitely becoming arbitrarily close to 0. He then drew the corresponding rectangles corresponding to the vertical lines through these points he would then have two collections of rectangles one inside and one outside the area in question. In our example above with $E = \frac{1}{2}$ this looks like:v



He then sums the areas of the corresponding rectangles: The inner being

$$(Em)^{\frac{p}{q}}(m - Em) + (E^2m)^{\frac{p}{q}}(Em - E^2m) + ... + (E^km)^{\frac{p}{q}}(E^{k-1}m - E^km) + ...$$

and the outer being

$$(m)^{\frac{p}{q}}(m - Em) + (Em)^{\frac{p}{q}}(Em - E^2 m) + ... + (E^{k-1}m)^{\frac{p}{q}}(E^{k-1}m - E^k m) + ...$$

The idea of Fermat is to add up the first k terms of these sums we first look at the outer sum and write it out

$$\begin{aligned} & m^{\frac{p}{q}+1} - Em^{\frac{p}{q}+1} + E^{\frac{p}{q}+1}m^{\frac{p}{q}+1} - E^{\frac{p}{q}+2}m^{\frac{p}{q}+1} + ... + E^{\frac{kp}{q}+k}m^{\frac{p}{q}+1} - E^{\frac{kp}{q}+k+1}m^{\frac{p}{q}+1} \\ = \; & m^{\frac{p}{q}+1}(1-E) + m^{\frac{p}{q}+1}(1-E)E^{\frac{p}{q}+1} + ... + m^{\frac{p}{q}+1}(1-E)E^{k\frac{p}{q}+k} + .. \\ = \; & m^{\frac{p}{q}+1}(1-E)(1 + E^{\frac{p}{q}+1} + ... + E^{k\frac{p}{q}+k} + ...). \end{aligned}$$

If we set $F = E^{\frac{p}{q}+1}$ then the outer sum is given as

$$m^{\frac{p}{q}+1}(1-E)(1 + F + F^2 + ... + F^k)$$

with $F = E^{\frac{p}{q}+1}$. Fermat writes this as $F = E^{\frac{p+q}{q}}$. We now recall that we can close this expression

$$1 + F + F^2 + ... + F^k = \frac{1 - F^{k+1}}{1 - F}.$$

He now has the expression

$$m^{\frac{p}{q}+1}\frac{(1-E)(1-F^{k+1})}{1-F}.$$

Now comes the brilliant "trick". We look at $G = E^{\frac{1}{q}}$ then $F = G^{p+q}$ and $E = G^q$. Thus we have

$$\frac{1-E}{1-F} = \frac{1-G^q}{1-G^{p+q}} = \frac{1-G^q}{1-G}\frac{1-G}{1-G^{p+q}}.$$

So

$$m^{\frac{p}{q}+1}\frac{(1-E)(1-F^{k+1})}{1-F} = m^{\frac{p}{q}+1}(1-F^{k+1})\frac{1-G^q}{1-G} \cdot \frac{1-G}{1-G^{p+q}}$$

which is equal to

$$m^{\frac{p}{q}+1}(1-F^{k+1})\frac{1 + G + ... + G^{q-1}}{1 + G + ... + G^{p+q-1}}.$$

Now the total sum over all values (i.e. not stopping at k) is larger than the indicated area. But the only part of the above expression that depends on k is the term $1 - F^{k+1}$. Which is always less than one. We conclude that for all values of E with $0 < E < 1$ the number

$$m^{\frac{p}{q}+1}\frac{1 + G + ... + G^{q-1}}{1 + G + ... + G^{p+q-1}}$$

is an upper bound for the area. If we evaluate this for $E = 1$ we get $m^{\frac{p}{q}+1}\frac{q}{p+q}$ as an upper bound for the area. If one looks at the expression for the sum of the inner rectangles it is just $E^{\frac{p}{q}}$ times the expression for the outer ones. We therefore find that

$$E^{\frac{p}{q}}m^{\frac{p}{q}+1}\frac{1+G+...+G^{q-1}}{1+G+...+G^{p+q-1}}$$

is a lower bound for the area. We can evaluate this at $E=1$ to see that the area is at least $m^{\frac{p}{q}+1}\frac{q}{p+q}$ and at most $m^{\frac{p}{q}+1}\frac{q}{p+q}$. Hence it must be equal to $m^{\frac{p}{q}+1}\frac{q}{p+q}$.

If we write $r=\frac{p}{q}$ then we have the familiar expression (for those who know some calculus) that the area is $\frac{m^{r+1}}{r+1}$.

Fermat used a similar method for negative powers, $r=-\frac{p}{q}$. Here one should look at the curve over the half line of all numbers $x>m$. The method involves taking $E>1$ and looking at the points $m<Em<E^2m<....$ This time the upper sum for the points $m,Em,...,E^km$ is:

$$m^{-\frac{p}{q}}(Em-m)+(Em)^{-\frac{p}{q}}(E^2m-Em)+(E^2m)^{-\frac{p}{q}}(E^3m-E^2m)+$$
$$...+(E^km)^{-\frac{p}{q}}(E^{k+1}m-E^km).$$

This time we can factor out $m^{-\frac{p}{q}+1}$ and have

$$m^{-\frac{p}{q}+1}(E{-}1{+}E^{-\frac{p}{q}+2}{-}E^{-\frac{p}{q}+1}{+}E^{-\frac{2p}{q}+3}{-}E^{-\frac{2p}{q}+2}{+}...{+}E^{-\frac{kp}{q}+k+1}{-}E^{-\frac{kp}{q}+k})=$$

$$m^{-\frac{p}{q}+1}(E-1)(1+E^{-\frac{p}{q}+1}+E^{-2\frac{p}{q}+2}+E^{-3\frac{p}{q}+3}+...+E^{-k\frac{p}{q}+k})=$$

$$m^{-\frac{p}{q}+1}(E-1)\frac{1-F^{k+1}}{1-F}$$

with $F=E^{1-\frac{p}{q}}=E^{-\frac{p-q}{q}}$. This time we write $G=E^{-\frac{1}{q}}$ and we have

$$m^{-\frac{p}{q}+1}E\frac{(1-G^q)(1-G^{(k+1)(p-q)})}{(1-G^{p-q})}.$$

Now if $p>q$ then as k is evaluated at increasinly large values the only term involving k is closer and closer to 1. As in the earlier case we have

$$m^{-\frac{p}{q}+1}E(1-G^{(k+1)(p-q)})\frac{1+G+...+G^{q-1}}{1+G+...+G^{p-q-1}}.$$

We see that the upper sum is always at most

$$m^{-\frac{p}{q}+1}E\frac{1+G+...+G^{q-1}}{1+G+...+G^{p-q-1}}.$$

Now eveluating at $E=1$ (thus $G=1$) we have as an upper bound on the area

$$m^{-\frac{p}{q}+1}\frac{q}{p-q}=\frac{m^{-\frac{p}{q}+1}}{\frac{p}{q}-1}.$$

Notice that this method only works for $\frac{p}{q} > 1$. If we need the area over a finite interval $0 < m < M$ then we can just subtract the area above M from the area above m and get

$$\frac{m^{-\frac{p}{q}+1}}{\frac{p}{q}-1} - \frac{M^{-\frac{p}{q}+1}}{\frac{p}{q}-1}.$$

If $0 < \frac{p}{q} < 1$ one can see that this formula is also true In the case of positive powers it is clear that if we wish an area for $0 < a < x < m$ then one can subtract the area between 0 and a and get

$$\frac{m^{r+1}}{r+1} - \frac{a^{r+1}}{r+1}$$

for $r = \frac{p}{q}$. We can see that the formula for the area over the same interval for $r = -\frac{p}{q} < 0$ but not -1 is

$$\frac{m^{r+1}}{r+1} - \frac{a^{r+1}}{r+1}.$$

Exercises.

1. For the indicated case of $y = x^{\frac{5}{3}}$, and $m = 2$ calculate (using a high precision calculator or math software package) the upper sums for $E = \frac{1}{2}, \frac{1}{3}, \frac{1}{5}$ and say 100 terms. Compare with the answer.
2. Why didn't the method above work for $r = -1$?
3. Complete the argument for the inner sum in the first part of the discussion.
4. Complete the argument for $r < -1$ by analyzing the lower sum.
5. What do you think Fermat did for rational numbers r with $0 > r > -1$?

4.2.3 Fermat's derivation of tangents.

In addition to his calculation of the area under the curves $y = x^r$ with r rational but not -1. Fermat also calculated the tangent lines. Here he also used methods that were clear precursors to what we call calculus. He observed that if one has a curve given as $y = x^n$ then the slope of the line through the points (x, x^n) and $(x + E, (x + E)^n)$ is

$$\frac{(x+E)^n - x^n}{(x+E) - x} = \frac{(x+E)^n - x^n}{E}.$$

The figure below is $y = x^3$ and A and B are two such points.

He observed that if E is chosen progressively smaller the connecting line would rotate to a tangent line (for the moment we will take this to mean that any line through A gotten by slightly rotating the tangent line intersects the curve at a nearby point, we will come back to the idea of a tangent line). Fermat (and probably many others) observed that if n is an integer then

$$(x+E)^n - x^n = nEx^{n-1} + \frac{n(n-1)}{2}E^2x^{n-2} + ... + E^n.$$

Thus every term is divisible by E. We therefore have

$$\frac{(x+E)^n - x^n}{E} = nx^{n-1} + \frac{n(n-1)}{2}Ex^{n-2} + ... + E^{n-1}.$$

He could then put $E = 0$ and finds the slope of the tangent line at (a, a^n) to be na^{n-1}. However, Fermat did more, he in fact calculated the slope of the tangent if n is only rational. Here we write $n = \frac{p}{q}$ and assume that $p, q > 0$. We are looking at

$$\frac{(x+E)^{\frac{p}{q}} - x^{\frac{p}{q}}}{E} = \frac{((x+E)^{\frac{1}{q}})^p - (x^{\frac{1}{q}})^p}{E}.$$

We note that if $E > 0$ then $(x+E)^{\frac{1}{q}} = x^{\frac{1}{q}} + F$ with $F > 0$. Thus taking q-th poweres of both sides of this equation we have

$$x + E = x + qFx^{1-\frac{1}{q}} + \frac{q(q-1)}{2}F^2x^{1-\frac{2}{q}} + ... + F^q.$$

Thus subtracting x form both sides of this equation and dividing by F we have

$$\frac{E}{F} = qx^{1-\frac{1}{q}} + \frac{q(q-1)}{2}Fx^{1-\frac{2}{q}} + ... + F^{q-1}.$$

This means that we can substitute E equals zero in this equation since if $E = 0$ then $F = 0$. We therefore have if $E = 0$ then we can evaluate $\frac{E}{F}$ and get $qx^{1-\frac{1}{q}}$. We now have

$$\frac{(x+E)^{\frac{p}{q}} - x^{\frac{p}{q}}}{E} = \frac{(x^{\frac{1}{q}} + F)^p - (x^{\frac{1}{q}})^p}{F}\frac{F}{E}$$

in both we can substitute $E = 0$ and get

$$p(x^{\frac{1}{q}})^{p-1}\frac{x^{\frac{1}{q}-1}}{q} = nx^{n-1}.$$

This certainly shows that Fermat knew a great deal of what we usually think of as basic calculus. However, he did not invent calculus. The point here is that by its very name calculus is a "method of computation". Fermat relies on brilliant relationships between rational and integral powers. He is not in the tradition of Archimedes either since he does not use true limits but rather uses a more algebraic formalism that allows substitution. We will discuss these distinctions more carefully when we get to our discussion of calculus.

4.2.4 Further precursors to calculus.

Mathematics flourished in the seventeenth century, Mathematicians finally had a notational system that had enough flexibility that they could study very general mathematical relationships. Also numbers had finally been divorced from units. Thus numbers could be manipulated algebraically without recourse (unless so desired) to geometric constructs. In Europe mathematicians were analysing areas, volumes, and tangents as they had never been before. As an example, we will take a look at the work of Isaac Barrow (1630-1663) who held the Lucasian Chair at Cambridge before Newton. He was more a geometer than an algebraist and had a low regard for abstract manipulation. His approach to the tangents studied by Fermat would be substantially as in the following discussion (we will, however, replace his geometric arguments with more algebraic ones). He would consider two positive relatively prime integers p and q yielding the curve that is the locus of points (x, y) with

$$y^q - x^p = 0.$$

To calculate the tangent to this curve at the point (a, b), fixed and on the courve, he would substitute $x = a + u$, $y = b + v$. Thus he would have

$$y^q - x^p = (b + v)^q - (a + u)^p =$$

$$b^q - a^p + qb^{q-1}v - pa^{p-1}u + E(u, v)$$

with $E(u, v)$ a sum of terms involving u^r or v^s with $r, s \geq 2$. The term $a^p - b^q = 0$ by assumption. Thus if u, v had been chosen very small and such that $(a + u, b + v)$ is on the curve then the quantity

$$qb^{q-1}v - pa^{p-1}u$$

must be very close to 0 (since if u is smaller than 1 then u^2 is smaller than u). This indicates that if we had a particle moving along the curve then at the point (a, b) it would be moving in the direction of the line

$$qb^{q-1}v - pa^{p-1}u = 0.$$

That is along the line

$$y = \frac{pa^{p-1}}{qb^{q-1}}x.$$

since $\frac{a^{p-1}}{b^{q-1}} = a^{\frac{p}{q}-1}$. This agrees with Fermat's solution. Barrow's approach is now called "implicit differentiation".

Exercises.

1. Complete the calculation that Barrow's method gives the same answer as Fermat's.

2. Use Barrow's method to calculate the tangent to the ellipse $x^2 + y^2 = 1$.

4.3 Calculus.

As we have seen, the first half of the seventeenth century was brimming with activity on calculations of areas and tangents. A substantial part of what we call calculus had already been discovered before either Newton or Leibniz had begun their work. However, it was these exceptional mathematicians who actually established the calculus. The term calculus means "a method of calculation". This is precisely what they developed. Their method unified what had been done before and established rules which if followed would lead to solutions to problems which heretofore were solved using ingeneious methods. As we have seen, one reason for the explosion of activity was the development of a notational system and an abstract formalism that simplified the task of communicating mathematics. In most aspects of the rivalry between Newton and Leibniz (actually the rivalry was between their adherents and desciples) the history gives the edge to Newton. However, when it comes to the notation that would be used in communicating and working with the calculus Leibniz wins hands down. Their independent work was published in several places. Leibniz published "A new method for maxima minima as well as tangents" in Acta Eruditorum, 1684. A year later Newton published "De methis Fluxionen" and claimed that the paper was written in 1671. Newton's masterpiece *Principia Mathematica* was published in 1687. In the introduction to the first edition he said:

"In letters that passed between me and that most excellent geometer G.W.Leibniz 10 years ago, when I signified that I knew a method of determining maxima and minima, of drawing tangents and the like, and when I concealed it in transposed letters... the most distinguished man wrote back that he had also fallen on a method of the same kind, and communicated his method which hardly differed from mine except in his forms of symbols."

The first calculus text was published in 1696 by the Marquis de L'Hospital called "Analyse des infinement petits" which was a compendium of lessons by his private tutor John Bernouli.

4.3.1 Newton's method of fluxions.

In this subsection we will describe Newton's approach to differential calculus. Since Leibniz' approach is essentially the same, we will emphasize the notational differences in the next subsection. Suppose we fix the independent variable to be x and y varies with x according to a predetermined rule. We think of the symbol o as indicating a very small change in x this symbol is a fluxion and at first we will take it to be an independently varying very small value. Then we think of $x + o$ as a very small change in x. Now when x has moved to $x + o$ the value of y changes to a new value $y + z$ (not Newton's notation). This z is an arbitrarily small change in y and so to Newton it should be proportional to o. That is $z = \dot{y}o$ and this proportionality should be a new function of x. The term $\dot{y}o$ is called a fluent and $\dot{y}$ is the derivative.

We will now look at an example. $y = x^n$ and n a positive integer. Then using the binomial formula we have

$$(x+o)^n = x^n + nx^{n-1}o + \text{higher powers of } o.$$

Since o is to be thought of as arbitrarily small the terms beyond the first power cannot contribute to the fluent. Thus the fluent if $nx^{n-1}o$ and $\dot{y} = nx^{n-1}$. This isn't much different from what Fermat might do. We nor look at the case when $n = \frac{p}{a}$.Then $y = x^{\frac{p}{q}}$ so $y^q = x^p$ (looks a bit like Barrow's start). This says that

$$(y + \dot{y}o)^q = (x+o)^p$$

Now

$$(x+o)^p = x^p + px^{p-1}o$$

and

$$(y+\dot{y}o)^q = y^q + qy^{q-1}\dot{y}o$$

thus equating coefficients of o we have

$$qy^{q-1}\dot{y}o = px^{p-1}o.$$

This implies that

$$\dot{y}o = \frac{p}{q}y^{1-q}x^{p-1}o = \frac{p}{q}x^{(1-q)\frac{p}{q}}x^{p-1}o = \frac{p}{q}x^{\frac{p}{q}-1}.$$

Notice that we have neglected the "higher powers of o". This is a consistant part of the method. The point here is that it is a method and not just a clever trick.

We look at one more example (which is a special case of the chain rule). Consider $y = \frac{1}{1-x}$. Then

$$(1-x)y = 1$$

so

$$(1-x-o)(y+\dot{y}o) = 1.$$

Expanding we have

$$(1-x)y - oy + (1-x)\dot{y}o = 1.$$

Using the relation we have

$$(1-x)\dot{y}o = oy$$

so

$$\dot{y}o = \frac{y}{1-x}o = \frac{1}{(1-x)^2}o$$

and we conclude

$$\dot{y} = \frac{1}{(1-x)^2}.$$

Exercises.

1. Calculate $\dot{y}$ for $y = x^3 + 3x + 1$ using the method of fluxions.

2. Suppose that you know $\dot{y}$ use the method of fluxions to calculate $\dot{z}$ if $z = \frac{1}{y}$.

3. Compare for the case of $y = x^n$ with n rational compare the method in this subsection with that of Fermat and that of Barrow.

4.3.2 The Leibniz notation.

The notation of Leibniz is now the standard method of expression in calculus. He wrote dx for Newton's $\dot{x}o$ and when $y = f(x)$ then what we denoted by $\dot{y}$ (this is not Newton's notation) he wrote $\frac{dy}{dx}$. In his notation one sees whar us called the chain rule in modern calculus immediately if $y = f(z)$ and $z = g(x)$ then

$$\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx}.$$

Leibniz is also credited with the product rule (also called the Leibniz rule). Suppose that $f(x) = g(x)h(x)$. That is $y = uv$ witt $u = g(x)$ and $v = h(x)$ then

$$\frac{dy}{dx} = \frac{du}{dx}v + u\frac{dv}{dx}.$$

In fact $y+dy = (u+du)(v+dv) = uv+vdu+udv+dudv = y+vdu+udv$. Now subtract y from both sides of the equation and divide by dx. This condition has come to be called Leibniz's rule.

Exercise.

1. Do problem 2. of the previous section using the chain rule.

4.3.3 Newton's binomial formula

Newton thought of o as small to first order, that is o^2 is negligable and he understood that one could equally well introduce objects small tosecond order say u with u, u^2 not negligable but. One would have

$$f(x+u) = f(x) + \dot{f}(x)u + \frac{1}{2}g(x)u^2$$

with $g(x)$ to be dertermined. He looked at $f(x) = x^{\frac{1}{m}}$ then $f'(x) = \frac{1}{m}x^{1-\frac{1}{m}}$. We now expand out

$$\begin{aligned} f(x+u)^m &= \left(f(x) + \dot{f}(x)u + \frac{1}{2}g(x)u^2\right)^m = \\ & f(x)^m + mf(x)^{m-1}\left(\dot{f}(x)u + \frac{1}{2}g(x)u^2\right) \\ & + \frac{m(m-1)}{2}f(x)^{m-2}(\dot{f}(x)u + \frac{1}{2}g(x)u^2)^2. \end{aligned}$$

Expanding in powers of u we have

$$\begin{aligned} x+u &= f(x+u)^m \\ &= x + mx^{\frac{m-1}{m}}(\frac{1}{m}x^{\frac{1-m}{m}}u + \frac{1}{2}g(x)u^2) + \frac{m(m-1)}{2}x^{\frac{m-2}{m}}\left(\frac{1}{m}x^{\frac{1-m}{m}}\right)^2 u^2 = \\ & x + u + \left(\frac{m}{2}x^{\frac{m-1}{m}}g(x) + \frac{m-1}{2m}x^{-1}\right)u^2. \end{aligned}$$

Thus

$$\frac{m}{2}x^{\frac{m-1}{m}}g(x)+\frac{m-1}{2m}x^{-1}=0$$

we can solve the equation and get $g(x)=-\frac{m-1}{m^2}x^{\frac{1}{m}-2}$. Observe that

$$-\frac{m-1}{m^2}=\frac{1}{m}(\frac{1}{m}-1).$$

So the third term is

$$\frac{\frac{1}{m}(\frac{1}{m}-1)}{2}x^{\frac{1}{m}-2}.$$

Proceding in this way Newton derived his formula that the $k+1$ term is

$$\frac{\frac{1}{m}(\frac{1}{m}-1)\cdots(\frac{1}{m}-k+1)}{k1}x^{\frac{1}{m}-k}$$

. Newton intruduced a notation analogous to ours for binomial coefficients to denote this expression. In modern notation we write

$$\binom{a}{k}=\frac{a(a-1)\cdots(a-k+1)}{k!}.$$

From this derivation he asserted that

$$(x+t)^{\frac{1}{m}}=x^{\frac{1}{m}}+\binom{\frac{1}{m}}{1}x^{\frac{1}{m}-1}t+\binom{\frac{1}{m}}{2}x^{\frac{1}{m}-2}t^2+...+\binom{\frac{1}{m}}{k}x^{\frac{1}{m}-k}t^k+...$$

This formally derived formula he checked by taking the m-th power.

One can then do the same for rational powers and get

$$(x+t)^a=x^a+\binom{a}{1}x^{a-1}t+\binom{a}{2}x^{a-2}t^2+...+\binom{a}{k}x^{a-k}t^k+...$$

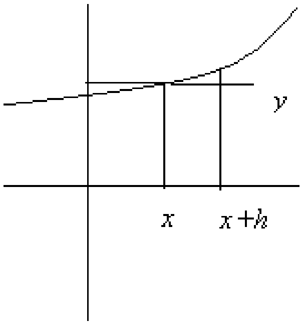
This is Newton's binomial series.

Exersizes.

1. Consider t or be the independent variable and calculate the derivative of $y=(x+t)^a$. Then differentiate the individual terms in Newton's series. Check that the two series for the same thing agree.

4.3.4 The fundamental theorem of calculus.

We consider rhe following picture



If we think of h as the infinitesimal o then we have the area $\dot{A}o$ under the curve above the interval is between yo and $(y+\dot{y}o)o$. But we can ignore the o^2 terms so $\dot{A}o = yo$. This says the $\dot{A} = y$. In other words the area under the curve $y = f(x)$ from a to x thought of as a function of x has derivative $f(x)$ at x. Thus the area under the curve between a and b with $a < b$ is $F(b) - F(a)$ for any function such that $\dot{F}(x) = f(x)$. This is the *fundamental theorem of calculus* which was first enunciated by Leibniz.

Of course, this argument is not in any way complete but it gives a method and that method yields the correct answer in all cases where another technique could be used. For example if $f(x) = x^m$ with $m \neq -1$ then $F(x) = \frac{1}{m+1}x^{m+1}$ and so we have the same outcome as Fermat (which was completely justified).

Exercises.

1. Use the fundamental theorem of calculus to derive the special case of Archimedes' theorem in exercise 2 of subsection 1.3 of this chapter.

2. This problem is difficult and can be considered a research project. Use the fundamental theorem of calculus to derive the theorem of Archimedes on the area of a sector of a parabola.

4.3.5 Logarithms

As we saw in the last subsection the Newton-Leibniz method has no problem calculation areas once a function is found with the appropriate derivative. The appropriate function for x^r is $\frac{1}{r+1}x^{r+1}$ except, of course, for $r = -1$. We will use the notation y' for what we wrote as $\dot{y}$ and $f'(x)$ for $\dot{f}$ (x). The question then remains what about $y' = \frac{1}{x}$? This is serious since it is necessity if we wish to calculate areas related to the hyperbola $uv = 1$. As it turns out the appropraite function had been discovered before calculus and for different reasons. We will digress from our main line and study the history of the "missing function". We first consider x as a function of y. Then $y(x(y)) = y$. Thus the chain rule says that $y'(x(y))x'(y) = 1$. But $y'(x(y)) = \frac{1}{x(y)}$. So

$$x'(y) = x(y).$$

In other words if we find a function such that $f'(x) = \frac{1}{x}$ then we would also find a function such that $g'(x) = g(x)$. Such a function with value $g(0) = 1$ has the remarkable property that $g(a+b) = g(a)g(b)$. That is, it changes addition into multiplication or vice-versa. We are ahead of our story.

John Napier (1550-1617) had an interest in making mechanical devices that would allow one to do complicated calculations precisely and easily. Before him there were several methods found of converting multiplication and division to addition and subtraction. One based on trigonometry that was perfected by the Arab mathematicians called prosthapaeresis. We will not go into this method here but suffice to say, it helped Tycho Brahe do his intricate calculations and was based on tables involving for sets of trigonometric identities involving multiplication addition and subtraction. Others, notably Michael Stifel (1487?-1567) had observed that if we fix a number a then

$$a^x a^y = a^{x+y}, \frac{a^x}{a^y} = a^{x-y}.$$

This certainly changes multiplication and division into addition and subtraction. However, for Stifel there were only caculations of such powers for a, x, y integers and for rational numbers one would need very accurate methods of extracting roots. Napier decided to just use integral powers of a number that had the property that the successive powers were close enough together that if one drew a straight line between the successive values (interpolated) the value would still be within a desired tolerance. This would allow one to use only integral powers in the tables or on the device to be constructed.

Napier chose the number $N = 0.9999999$. He then considered N^L to have logarithm L. In order to avoid small decimals he in fact considered $10000000N^L$ as having logarithm L.Now to multiply $10000000N^L$ by $10000000N^K$ all you need do is add $K + L$ look at the table to find the number with logarithm $K + L$ (or interpolate to get it) and then shift by 7 decimal positions. This was implemented in a slide rule type mechanism. Note that if $L = 1$ corresponds to the number 9999999 and $L = 0$ to 10000000. Now if we calculate he value $10000000N^L$ for $L = 10000000$ we get 3678794 to seven digit accuracy. It therefore gave an efficient method of doing 7 digit multiplication and division. But except for turniing multiplication into division what does it have to do with the problem of finding a function whose derivative is $\frac{1}{x}$?

A hint can befound in the following observation. If $f(x)$ satisfyies $f'(x) = f(x)$ and $f(0) = 1$ then to seven significatnt figures $f(1) = 2.718281$. The reciprical of this number is 0.3678794 to 7 decimal places. This cannot be an accident.

The upshot is that a function whose derivative is $\frac{1}{x}$ is very different then a function whose derivative is x^n for any integer other than -1. The function that has this derivative and value 0 at 1 is usually denoted $\ln(x)$ and is called the natural logarithm. It is also denoted simply as $\log(x)$ when logarithms to base 10 are note being used. Convarting a base involves the simple maneuver of multiplying by the logarithm of the inverse. Thus $\log(x) = \frac{\ln(x)}{\ln(10)}$. The number that yields a natural logarithm of 1 is usually denoted by e. This number is not rational and as observed above it is 2.718281 to seven decimal places.

Exercise.

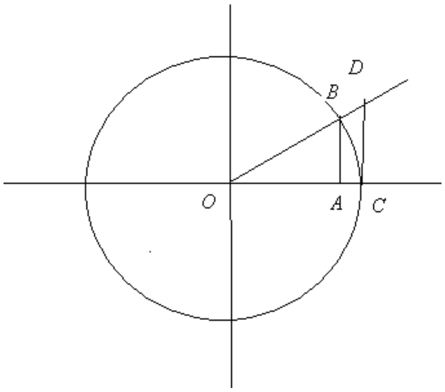
Suppose Napier had used $100000000 = 10^8$ so he would have been looking at powers of $N = .99999999$. What would the Napier logarithm of $10^8 N^{10^8}$ be?

4.3.6 The trigonometric functions.

We have seen that the ancient Greeks had an extensive knowledge of trigonometry. We have given an interpretation of trigonometry in section 3.5.3. In particular, the two basic trigonometric functions $\cos(x)$ and $\sin(x)$. Notice that we are using x for the variable rather than a more traditional Greek letter and thinking of the functions as being attributes of angles. We have seen that these functions have the following properties:

1. $\cos(0) = 1, \sin(0) = 0$.
2. $\cos(x)^2 + \sin(x)^2 = 1$.
3. $\sin(x + y) = \sin(x)\cos(y) + \sin(y)\cos(x)$.
4. $\cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y)$.

We will calculate the derivatives of these functions using techniques of analysis. We will use the prime rather than the dot notation. Consider the picture below of a circle of radius 1.



The lengh of AB is the $\sin(\theta)$ where θ is the angle AOB. The length of OA is $\cos(\theta)$. The length of the arc CB is θ and the length of CD is $\tan(\theta) = \frac{\sin(\theta)}{\cos(\theta)}$. We note that the area of the triangle COB is $\frac{\sin(\theta)}{2}$. The area of the triangle COD is $\frac{\tan(\theta)}{2}$ (see exercise 1below) and the area of the part of the interior of the circle COB is $\frac{\theta}{2}$(at least for θ positive and a rational multiple of $\frac{\pi}{2}$ that is at most $\frac{\pi}{2}$). To see this last assertion note that the area of a quadrant is $\frac{\pi}{4}$ (since the area of the interior of the circle is π). To get a quadrant we take $\theta = \frac{\pi}{2}$. If $\theta = \frac{\pi}{4}$ then we would get half the area which is $\frac{\pi}{8} = \frac{\theta}{2}$. If $\theta = \frac{\pi}{2k}$ with k a positive integer than we would have area $\frac{1}{k}$ times that of the area of the quadrant. That is $\frac{\theta}{2}$. If we multiply θ by a positive integer m and θ is very small then we get the area of m equal pieces corresponding to θ. Thus the area is $\frac{m\theta}{2}$. We therefore see that if θ is a positive rational multiple of $\frac{\pi}{2}$ less than or equal to $\frac{\pi}{2}$ then the area of the piece of the circle is $\frac{\theta}{2}$. We now observe that since the three areas are nested we have

$$\frac{\sin(\theta)}{2} < \frac{\theta}{2} < \frac{\sin(\theta)}{2\cos(\theta)}$$

in the range $0 < \theta < \frac{\pi}{2}$. We therefore see that $\frac{\sin(\theta)}{\theta} < 1$ and $\frac{\sin(\theta)}{\theta} > \cos(\theta) = \sqrt{1 - \sin(\theta)^2}$. Using $0 < \sin(\theta) < \theta$ we see that $\frac{\sin(\theta)}{\theta} > \sqrt{1 - \theta^2}$. From this we see that as we choose θ positive and progressively smaller the value of $\frac{\sin(\theta)}{\theta}$ is being crushed to 1. This says that the slope of the tangent line to $y = \sin(x)$ at $x = 0$ is 1.That is $\sin'(0) = 1$. It is easier to see that $\cos'(0) = 0$. Indeed, since $\cos(x)^2 + \sin(x)^2 = 1$ we can use Leibniz rule to see that

$$2\cos'(x)\cos(x) + 2\sin'(x)\sin(x) = 0.$$

Subsituting $x = 0$ gives

$$2\cos'(0)\cos(0) + 2\sin'(0)\sin(0) = 0.$$

We have $\cos(0) = 1$ and $\sin(0) = 0$. So $\cos'(0)$ is indeed 0.

To calculate all derivatives we note $\sin(x+o) = \sin(x)\cos(o) + \cos(x)\sin(o)$. Now $\cos(o) = \cos'(0)o = 0$ and $\sin(o) = \sin'(0)o = o$. So $\sin(x+o) = \cos(x)o$. Similarly, $\cos(x+o) = \cos(x)\cos(o) - \sin(x)\sin(o) = -\sin(x)o$.

This yields

4. $\cos'(x) = -\sin(x)$ and $\sin'(x) = \cos(x)$.

In the above derivation we used the fact that by choosing θ is small we can make $\sqrt{1-\theta^2}$ as close to 1 as we wish. This is true and you might think that it is obvious but it does need proof in modern mathematics. We will discuss this point in the exercises.

Exersizes,

1. Use the theory of similar triangles to deduce that in the figure above $CD = \tan\theta$, (Hint: $\frac{CD}{AB} = \frac{OC}{OA}$.

2. Here we will sketch that assertion that if $-1 \leq \theta \leq 1$ then $1 - \sqrt{1-\theta^2} < \theta^2$. First check that

$$\theta^2 = (1 - \sqrt{1-\theta^2})(1 + \sqrt{1-\theta^2}).$$

Next observe that in the range indicated $1 + \sqrt{1-\theta^2} \geq 1$. Conclude

$$1 - \sqrt{1-\theta^2} = \frac{\theta^2}{(1 + \sqrt{1-\theta^2})} \leq \theta^2.$$

4.3.7 The exponential function.

In this section we will discuss Euler's unification of logarithms and trigonometry. This was essential done section 3.5.3. We will first take another look at logarithms. We saw that in Napier's work on logarithms the number

$$u = (1 - \frac{1}{n})^n$$

with $n = 10000000 = 10^7$ played an important role. Also, we pointed out that if $f(x)$ were a function with $f'(x) = f(x)$ and $f(0) = 1$ then to seven significant figures $f(1) = \frac{1}{u}$. Euler established the standard notation $e = f(1)$. Now, $f(x)$ satisfies $f(x+y) = f(x)f(y)$. This is reminicent of the known formula $a^{x+y} =$

$a^x a^y$ for x, y rational. Further, $a^0 = 1$ is the standard interpretation of the 0-th power and we have by definition $a^1 = a$. This led to the notation $f(x) = e^x$. The distinction is that this function is defined for all real numbers. If we have $a = e^L$ then we say that $L = \ln(a)$. This defines the natural logarithm that is up to a sign and a shift essentially Napier's logarithm (that is to 7 significant figures). We have seen that $\ln'(x) = \frac{1}{x}$ giving the "missing" derivative.

We note that this new function allows us to define a^x for $a > 0$ and any real number, x, by $a^x = e^{\ln(a)x}$.

In section 3.5.3 we also saw that in the realm of complex numbers if we set

$$z(\theta) = \cos(\theta) + i\sin(\theta)$$

then the trigonometric identities of the previous section can be written

$$z(\alpha + \beta) = z(\alpha)z(\beta).$$

We also note that $z(0) = 1$. This led Euler to define $e^{ix} = \cos(x) + i\sin(x)$.

This allowed the exponential function to be defined for all complex numbers as

$$e^{x+iy} = e^x \left(\cos(y) + i\sin(y)\right).$$

The basic properties are still satisfied in this context.

1.$e^0 = 1$.
2.$e^{z+w} = e^z e^w$.

Euler was especially intrigued with the formula

$$e^{i\pi} + 1 = 0$$

which he called the relationship between the 5 most important constants of mathematics.

At this point we are ahead of our story. We need to learn a few things from Euler's teachers.

Exercises.

1. What are the 5 constants in Euler's formula?

2. What value would you assign to $(e^{i\pi})^{i\pi}$? How would you interpret the value?

4.3.8 Power series expansions.

We have already encountered Newton's bininomial formula which is an infinite series. This formula showed how one might express a function as an infinite series. In this case it is the function $\frac{1}{(1-x)^a}$ with a rational and $-1 < x < 1$. One notes that if we differentiate this k times one gets

$$\frac{a(a+1)\cdots(a+k-1)}{(1-x)^{k+a}}.$$

Thus if we call the function $f(x)$ then Newton's binomial formula becomes

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + ... + \frac{f^{(k)}(0)}{k!}x^k + ...$$

Here $f^{(k)}$ is gotten by differentiating f repeatedly k times. This result was generalized by many authors but most notably Brook Taylor (1685-1773) and later Colin Maclaurin(1698-1746) who are interchangeably named for the generalization. It says that a function can be expanded in the form

$$f(c + x) = f(c) + f'(c)(x - c) + \frac{f''(c)}{2}(x - c)^2 + ... + \frac{f^{(k)}(c)}{k!}(x - c)^k + ...$$

We will call this series the Taylor series of $f(x)$ at c.

For example of $f(x) = e^x$. Then $f'(x) = f(x)$ and $f(0) = 1$. Thus $f''(x) = (f')'(x) = f'(x) = f(x)$. So $f^{(k)}(0) = 1$ for all k. This says that the Taylor series of e^x is

$$1 + x + \frac{x^2}{2} + \frac{x^3}{6} + ... + \frac{x^k}{k!} + ...$$

Similarly we have

$$\cos'(x) = -\sin(x) \text{ and } \sin'(x) = \cos(x).$$

This gives

$$\cos''(x) = -\sin'(x) = -\cos(x).$$

and

$$\sin''(x) = \cos'(x) = -\sin(x).$$

This says that even repeated derivatives are given as follows

$$\cos^{(2k)}(x) = (-1)^k \cos(x) \text{ and } \sin^{2k}(x) = (-1)^k \sin(x).$$

The odd repeated derivatives are given as

$$\cos^{2k+1}(x) = (-1)^k \cos'(x) = (-1)^{k+1} \sin(x)$$

and

$$\sin^{(2k+1)}(x) = (-1)^k \sin'(x) = (-1)^k \cos(x).$$

Since $\cos(0) = 1$ and $\sin(0) = 0$ we have the Taylor series

$$\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{4!} - ... + (-1)^k \frac{x^{2k}}{(2k)!} + ...$$

and

$$\sin(x) = x - \frac{x^3}{6} + \frac{x5}{5!} - ... + (-1)^k \frac{x^{2k+1}}{(2k+1)!} + ...$$

If we add together $\cos(x) + i\sin(x)$ then we have

$$1 + ix - \frac{x^2}{2} - \frac{ix^3}{3!} + \frac{x^4}{4!} + i\frac{x^5}{5!} + ... + (-1)^k \frac{x^{2k}}{(2k)!} + (-1)^k \frac{x^{2k+1}}{(2k+1)!} + ...$$

if we write $z = ix$ then $z^{2k} = (i^{2k})x^{2k} = (i^2)^k x^{2k} = (-1)^k x^{2k}$ and $z^{2k+1} = ix(z^{2k}) = ix(-1)^k x^{2k} = (-1)^k i x^{2k+1}$. Thus in terms of z we have

$$1 + z + \frac{z^2}{2} + \frac{z^3}{3!} + ... + \frac{z^k}{k!} + ...$$

This says that Euler's interpretation of e^{ix} as $\cos(x) + i\sin(x)$ is completely consistant with Taylor series.

4.3.9 Euler's summation of a series.

As we mentioned the value of the series

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + ... + \frac{1}{n^2} + ...$$

was a mystery to some of the greatest minds of the seventeenth and early eighteenth centuries. We have discussed the method of Leibniz that summed the series

$$1 + \frac{1}{3} + \frac{2}{3 \cdot 4} + ... + \frac{2}{n(n+1)} + ... = 2.$$

One notes that each term of the first series is less than the corresponding one in the second. This implies that the series sums to a number that is infact less than 2.

Before we give Euler's ingeneous deduction of the sum there is another sum that the previous section allows us to calculate.

$$1 + \frac{1}{2} + \frac{1}{6} + ... + \frac{1}{n!} + ... = e.$$

The terms in the sum are fairly simple but the number e is not a simple rational number. One doesn't guess such a value and in fact it had no name until Euler named it. It is therefore not a reasonable idea to just guess an answer.

He first observes that if we have a polynomial of the form

$$1 - a_1 x + a_2 x^2 + ... + a_n x^n$$

and if this polynomial has roots $r_1, ..., r_n$ counting multiplicity then if this roots are non-zero we have

$$a_1 = \frac{1}{r_1} + \frac{1}{r_2} + ... + \frac{1}{r_n}.$$

To see this we observe that the polynomial with value 1 at 0 and roots $r_1, ..., r_n$ is

$$\left(1 - \frac{x}{r_1}\right)\left(1 - \frac{x}{r_2}\right) \cdots \left(1 - \frac{x}{r_n}\right).$$

Now compare the coefficient of x. We will come back to this in the exercises. Euler's leap was to apply this observation to an infinite series (as in the last section). He considers

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - ... + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + ...$$

Thus

$$\frac{\sin(x)}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - ... + (-1)^n \frac{x^{2n}}{(2n+1)!} + ...$$

as series with only even powers. Assuming that we can expand it as a polynomial the roots being $\pm n\pi$ wint $n = 1, 2, 3, ...$ So we could expect that this is given by

$$\begin{aligned} & \left(1 - \frac{x}{\pi}\right)(1 + \frac{x}{\pi})\left(1 - \frac{x}{2\pi}\right)(1 + \frac{x}{2\pi})\cdots\left(1 - \frac{x}{n\pi}\right)(1 + \frac{x}{n\pi})\cdots \\ = \quad & \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{4\pi^2}\right)\cdots\left(1 - \frac{x^2}{n^2\pi^2}\right)\cdots \end{aligned}$$

which also has even powers. Thus if you consider the series

$$1 - \frac{x}{3!} + \frac{x^2}{5!} - ... + (-1)^n \frac{x^n}{(2n+1)!} + ...$$

then it is reasonable to think that it is given by

$$\left(1 - \frac{x}{\pi^2}\right)\left(1 - \frac{x}{4\pi^2}\right)\cdots\left(1 - \frac{x}{n^2\pi^2}\right)\cdots$$

Thus if we apply the observation (valid for polynomials) we have

$$\frac{1}{3!} = \frac{1}{\pi^2} + \frac{1}{4\pi^2} + ... + \frac{1}{n^2\pi^2} + ...$$

This yields

$$\frac{\pi^2}{6} = 1 + \frac{1}{4} + \frac{1}{9} + ... + \frac{1}{n^2} + ...$$

Although no one doubted this as the sum of the series after they saw the marvelous argument the reader should be cautioned that this is not a proof of the formula (as Leibniz' derivation is of his value for his series). As it turns out this argument can be made rigorous using a theory of infinite products. Indeed, the above infinite products can be be proved to converge in a well defined sense to the desired function and the suggested formal manipulation actually gives the Taylor series.

Exercise.

1. If f is a polynomial of degree n with roots $r_1, ..., r_n$ (allowing for repititions) then f is a multiple of

$$(x - r_1)...(x - r_n).$$

Assuming that $f(0) = 1$ show that

$$f(x) = \left(1 - \frac{x}{r_1}\right)\left(1 - \frac{x}{r_2}\right)\cdots\left(1 - \frac{x}{r_n}\right).$$

Hint: The multiple is $\frac{(-1)^n}{r_1 r_2 \cdots r_n}$.

4.3.10 The question of rigor.

There were two controversies that arose in the development of the Calculus. The first was the question of priority between Newton and Leibniz. It can be said that Newton came out ahead on that issue (although few doubt the independence of Leibniz' contribution). However, Newton's apparent victory was one of the causes of the eclipse of English mathematics during the eighteenth century. There are many explanations of this but one the strongest (to our mind) is just that the Leibniz notation was superior. The second contraversy had to do with the very roots of the Calculus.

The scientific community knew that caculus gave them an entirely new arsonal of tools to study problems in simple mechanical ways that had been only handled in special cases by methods that were extremely clever and complicated. However, the method of both Newton and Leibniz involved the multiplication and division of objects that were not exactly numbers. Newton's symbol o was an object that one should consider to be such that o^2 can be neglected in expressions where it occurs.. The ratio $\frac{f(x+o)-f(x)}{o} = f'(x)$ is the derivative. Leibniz' approach was similar he had dx and one should think of $(dx)^2 = 0$ but $\frac{dy}{dx}$ was the derivative. Many scientists, philosophers,etc. felt that there was a dangerous lack of foundation for these methods. However, the methods always gave correct answers to the problems to which they were applied.

However, the application of the methodology was becoming more and more of a specialty. For example, in the derivation that we gave in the previous section we saw an argument that as it turns out gives the correct answer but is based on a premise that has not been checked. One starts with (at least the hope that) something like the following statement is true.

1. $f(0) = 1$.

2. $f'(0) = -a$.

Then a is the sum of the reciprocals of the roots of f (the numbers c such that $f(c) = 0$.

This is true for polynomials if one includes complex roots. However it is definitely false for even very nice functions. For example, e^x has the properties 1. and 2. with $a = -1$. But it is never 0 (this includes using the extended definition $e^{x+iy} = e^x(\cos y + i \sin y)$ of Euler. This says that the argument of Euler is only rigorous if he shows that the function that he defined has the property that the sum of the reciprocals of its roots is the negative of its derivative at 0.This can be done, as we indicated in the previous section by giving a rigorous meaning to the product formula for $\frac{\sin x}{x}$.

Even before Euler at the very beginning of the development of Calculus there were skeptics about the foundations (not the applications). One of the most serious attacks was made by Bishop George Berkeley (1685-1753). In his pamphlet *The Analyst* in 1734 he expressed doubts about the foundations of Calculus in particular of Newton's fluxions. His point was you cannot have something that behaves like a very small but non-zero number and still has the property that its square is 0. It is clear that you must exercise great care when yo divide by something whose square is o. He labled such objects infinitesimals and argued that they cannot have an independent reality. Here is a quote from *The Analyst* in his discussion of fluxions:

...they are neither finite quantities nor quantities infinitely small, not yet nothing.

In fact, one can develop a rigorous theory with highly restricted classes of functions. For example, if we only consider polynomials then we well see in the next chapter that we can have a completely consistant theory with polynomials in two variables with one variable having the property that it is not 0 but its square is 0. This theory would also wallow for power series and explain why the Newton-Leibniz method always gave the right answer for functions given by power series. A more radical consistant theory which allowed for objects like the ones that Berkeley disparaged was developed by Abraham Robinson in his theory of non-standard analysis. Roughly speaking, he hypthesized the existance of non-standard numbers that were allowed to “fit between” actual numbers. To describe these numbers we must understand our usual number system in a more rigorous manner than we have so far.

There were other problems with the foundations of calculus that were less apparent in the seventeenth and eighteenth centuries. This had to do with how careful one must be in the choice of functions that are analyzable using the methods at hand. For example if we considered the function $|x|$ (x if $x \geq 0$ and $-x$ if $x < 0$) Then if $x > 0$ we have

$$\frac{|x+o|-|x|}{o} = \frac{x+o-x}{o} = 1$$

and if $x < 0$ we have

$$\frac{|x+o|-|x|}{o} = \frac{-(x+o)-(-x)}{o} = \frac{-o}{o} = -1.$$

However if we consider $x = 0$ then we are dealing with

$$\frac{|o|}{o}.$$

In other words we must figure out a meaning for $|o|$. much worse phenomena are possible and can actually occur in useful applications of mathematics. Calculus had to be given a firm footing,