Contents of MTH1001–Algebra

Sandro Mattarei, University of Lincoln

Lecture notes of Algebra. Week 3

7. Writing numbers in a different base

7.1. Integers, and then real numbers, written in a base b. Fix an integer $b > 1$, called *base*. Then every non-negative integer n can be written in the form

$$n = d_{k-1}b^{k-1} + d_{k-2}b^{k-2} + \cdots + d_1 b + d_0,$$

and will be denoted by $(d_{k-1} \dots d_1 d_0)_b$, where each d_i is a digit in base b, that is, a symbol for one of the integers $0, 1, \dots, b-2, b-1$. For example, if $b \leq 10$ one one can use the ordinary (decimal) $0, 1, \dots, b-1$ digits to represent themselves in base b. However, if $b > 10$ it may be convenient to use extra symbols to denote the digits having values $10, 11, \dots, b-1$. For example, when $b = 16$ (the *hexadecimal* system) it is customary to use the letters from A to F as digits with values 10 to 15.

It is neither necessary nor convenient to assume $d_{k-1} \neq 0$. If that holds then we say that n has (exactly) k digits in base b. Not assuming that allows us to identify writings such as 010 or 0010 for the number 10 in decimal notation. Allowing for that we have that the above expression for n in base b is unique, that is, each digit d_j (including the leading zeroes) depends only on n. [3]

As for base 10, this generalises from positive integers to arbitrary positive real numbers. Every positive real number n can be written as $\sum_{i<k} d_i b^i$ (with i ranging over all integers less than k, hence possibly negative) and denoted by $(d_{k-1}d_{k-2} \dots d_1 d_0, d_{-1}d_{-2} \dots)_b$, where the digits may continue indefinitely to the right. This writing is also unique, but with exceptions, occurring when all d_i starting with a certain d_j equal $b-1$; in that case we may replace d_i for $i \geq j$ with 0 (and possibly omitting it in writing) and increase d_{j-1} by one.

Remark. The number of digits in base b of a non-negative integer n is given by the formula

$$k = \lfloor \log_b n \rfloor + 1 = \left\lfloor \frac{\log n}{\log b} \right\rfloor + 1.$$

This is because $b^{k-1} \leq n < b^k$.

[3] A proof of uniqueness essentially amounts to writing up formally and more generally the familiar way in which we tell which of two different integers in base 10 is larger, by scanning the digits from left to right until we find a different digit (assuming they have the same number of digits). A proof of existence of the expression, for any positive integer n, can be obtained by writing up formally the procedure for writing and integer in an arbitrary base b, which we explain later.

7.2. Converting integers from base b to base 10. For n an integer we conveniently write the conversion as

$$n = (\dots((d_{k-1}b + d_{k-2})b + d_{k-3})b + \cdots + d_1)b + d_0.$$

This method (the same as we conveniently use to evaluate a polynomial on some number, in thic case the base b) requires only $k-1$ multiplications by b, and $k-1$ additions. It is also easy to perform on a pocket calculator (as long as that is not too smart, meaning that it should execute operations in the order in which we type them in.) For example,

$$(61405)_7 = ((6 \cdot 7 + 1) \cdot 7 + 4) \cdot 7 \cdot 7 + 5 = 14950.$$

The calculations can be conveniently arranged as follows, a special case of Ruffini's rule for polynomials (or Horner scheme, see a later section on polynomials):

	6	1	4	0	5
7		42	301	2135	14945
	6	43	305	2135	14950

7.3. Converting integers from base 10 to base b. If n is an integer, its last digit in base b, which is d_0, can be obtained as the remainder of dividing n by b, then d_1 is obtained as the remainder of dividing the previous quotient by b, etc., until we get quotient zero. For example converting $(14950)_{10}$ to base 7 can be done as follows:

$$\begin{aligned} 14950 &= 7 \cdot 2135 + 5 \\ 2135 &= 7 \cdot 305 + 0 \\ 305 &= 7 \cdot 43 + 4 \\ 43 &= 7 \cdot 6 + 1 \\ 6 &= 7 \cdot 0 + 6 \end{aligned}$$

We conclude that $(14950)_{10} = (61405)_7$.

This algorithm can be efficiently executed on a pocket calculator (which does not do divisions with remainder) if we repeatedly divide by b starting with n (without subtracting the remainders), and comparing the fractional part of each quotient with a table which we will have prepared in advance, containing $0/b, 1/b, 2/b, \dots, (b-1)/b$. The correct digit d at each stage will be found according to the rule $d/b \leq f < (d+1)/b$, where f is the

fractional part of the quotient. In the previous example we will have

$$\begin{aligned} 1/7&=0.142\ldots\\ 2/7&=0.285\ldots\\ 3/7&=0.428\ldots\\ 4/7&=0.571\ldots\\ 5/7&=0.714\ldots\\ 6/7&=0.857\ldots \end{aligned}$$

We will find

$$\begin{array}{r} 14950\quad . \\ 2135.714\ldots \\ 305.102\ldots \\ 43.586\ldots \\ 6.226\ldots \\ 0.889\ldots \end{array} \quad \text{and hence} \quad \begin{array}{|c|c|} \hline 0.714 & 5 \\ \hline 0.102 & 0 \\ \hline 0.586 & 4 \\ \hline 0.226 & 1 \\ \hline 0.889 & 6 \\ \hline \end{array}$$

Note that such a table will need to be sufficiently accurate, especially if there are sequences of consecutive digits 6 in the expansion of n in base b. Also, the divisions will need to be done with a sufficient accuracy, otherwise we may find an incorrect answer, as we exemplify now.

Example. Consider the integer 16806, in decimal notation. If we perform the divisions by 7 with a precision limited to three digits after the point, we find

$$\begin{array}{r} 16806\quad . \\ 2400.857\ldots \\ 342.980\ldots \\ 48.997\ldots \\ 7.000\ldots \\ 1\ldots \\ 0.142\ldots \end{array} \quad \text{and hence} \quad \begin{array}{|r|c|} \hline 0.857 & 6 \\ \hline 0.980 & 6 \\ \hline 0.997 & 6 \\ \hline 0.0 & 0 \\ \hline 0.0 & 0 \\ \hline 0.142 & 1 \\ \hline \end{array}$$

It would appear that $(16806)_{10} = (100666)_7$, but that is wrong, and the correct conclusion is $(16806)_{10} = (66666)_7$. The problem is that

$$\frac{6}{7} + \frac{6}{7^2} + \frac{6}{7^3} + \frac{6}{7^4} = 0.99958307\ldots,$$

which gets approximated to 1 if we only use three digits after the point (In other words, the 7.000 appearing in the penultimate row of the above calculation should really be a little less than 7. Unfortunately, this may produce a large error in the final result: in this case an error of $(1000)_7 = 7^3$.)

REMARK. We have used two different methods to convert from base b to base 10, and from base 10 to base b (inverse to each other: *multiplying* in the former case, and *dividing* in the latter). By symmetry, each of those methods would also work for the other task, but it would involve doing the relevant calculations in base b rather than in base 10. Of course if $b = 2$ and the conversions are to be done by a computer working in base 2, rather than by a human, the two algorithms would be exchanged.

REMARK. Converting from *binary* (base 2) to *hexadecimal* (base 16, where the customary symbols for the digits are $0, 1, \ldots, 9, A, B, C, D, E, F$), and from hexadecimal to binary, will be much simpler than described above. For example, to convert from binary to hexadecimal it will be sufficient to split the bits into blocks of four starting from the decimal points, and then convert each block into the corresponding hexadecimal digit.

7.4. Converting real numbers from base b to base 10. Consider a positive number written in base b with a finite number h of digits after the point. To convert it to decimal one may use the same procedure as for an integer, but continuing with the digits (in base b) after the point, up to the last digit d_{-h}, and then divide the result by b^h. (This is because ignoring the point amounts to multiplying our number by b^h.)

EXAMPLE. To convert $(14.22)_5$ to decimal, remove the point (which means multiplying by 5^2), convert $(1422)_5 = 237$, and then divide by 5^2: $(14.22)_5 = 237/25 = 9.48$.

Note that until just before this last step you work with integer numbers, thus avoiding approximation errors. In fact, because of the final division the decimal expansion of the number may have infinitely many digits, even though you started with finitely many digits in base b. This cannot occur if $b = 2$ or 5 or, more generally, if the base b has only 2 and 5 as prime factors. (See a later subsection about periodic numbers.)

EXAMPLE. We have $(1.22)_3 = (122)_3/3^2 = 17/9 = 5.222\cdots = 5.\dot{2}$.

If the number to be converted has infinitely many digits, one can do the same with an approximation (keeping a few digits after the point).

EXAMPLE. To convert $(2.\dot{1})_3 = (2.11111\cdots)_3$ into decimal, we may convert the approximation $(2.111)_3 = (2111)_3/27 = 67/27$, which equals $2 + \frac{13}{27} = 2.\dot{4}8\dot{1}$, so roughly 2.48. In this case we can actually convert it exactly, namely $(2.\dot{1})_3 = 2.5$, because $(2.\dot{1})_3 \cdot 2 = (11.\dot{2})_3 = (12)_3 = 5$. More generally, if the number is periodic (written in any base) then it is rational and there is a rule to convert it into a fraction of integers, see a later subsection.

7.5. Converting real numbers from base 10 to base b. If n is not an integer we can deal separately with the integer part and the fractional part. In fact, multiplying

the latter by b and taking the integer part of the result we get d_{-1}; then we may repeat this procedure with the fractional part of the result and find d_{-2}, etc.

EXAMPLE. To convert 2.481 to base 3, write it as $2 + 0.481$.

- $0.481 \cdot 3 = 1.443$, so first digit after the point will be 1;
- $0.443 \cdot 3 = 1.329$, so second digit after the point will be 1;
- $0.329 \cdot 3 = 0.987$, so third digit after the point will be 0;
- $0.987 \cdot 3 = 2.961$, so fourth digit after the point will be 2;
- $0.961 \cdot 3 = 2.883$, so fifth digit after the point will be 2; and so on.
- In conclusion, $2.481 = (2.11022\cdots)_3$.

This can be conveniently done on a simple pocket calculator, and we can also avoid subtracting the integral part at each step by using a trick which we have seen before, and requires a preliminary table with (at least approximate) decimal expansions of $1/b, 2/b, \ldots (b-1)/b$. In this case we have $1/3 = 0.\dot{3}$ and $2/3 = 0.\dot{6}$. We proceed as before, but reading each digit off the fractional part (rather than from the integral part) *before* multiplying by b, hence subtracting the integral part becomes unnecessary.

Hence to convert 2.481 to base 3, write it as $2 + 0.481$. Then

- because $\dot{3} \leq 0.481 < \dot{6}$, the first digit after the point will be 1;
- $0.481 \cdot 3 = 1.443$, and as $\dot{3} \leq 0.443 < \dot{6}$ the second digit after the point will be 1;
- $1.443 \cdot 3 = 4.329$, and as $0.329 < \dot{3}$ the third digit after the point will be 0;
- $4.329 \cdot 3 = 12.987$, and as $\dot{6} \leq 0.987$ the fourth digit after the point will be 2;
- $12.987 \cdot 3 = 38.961$, and as $\dot{6} \leq 0.961$ the fifth digit after the point will be 2;
- $38.961 \cdot 3 = 116.883$, and as $\dot{6} \leq 0.883$ the sixth digit after the point will be 2; and so on.
- In conclusion, we find $2.481 = (2.110222\cdots)_3$.

EXAMPLE. Converting the number π to base 2 we will find:

$$(3,1415926\ldots)_{10} = (11,00100100001111110\ldots)_2.$$

Please check that yourself, starting with as many decimal digits of π as they fit on your calculator, and proceeding like in the previous example, comparing the fractional part with $1/2 = 0.5$ after each multiplication by 2.

8. Arithmetic and geometric progressions

A finite sequence $a_1, a_2, \ldots, a_n$ of (real or) complex numbers is called an *arithmetic progression* if the difference $d = a_{k+1} - a_k$ between each two consecutive terms is constant, meaning that it is independent of k. We could also write this condition as $a_{k+1} - a_k = a_k - a_{k-1}$ for all k (to which this applies, that is, $1 < k < n$), or, equivalently, $a_{k+1} + a_{k-1} = 2a_k$.

Hence a sequence is an arithmetic progression precisely when each term is the average, or arithmetic mean, of the preceding and the following term. [4]

For an arithmetic progression we have

$$a_n = a_1 + d(n-1).$$

The sum of an arithmetic progression, also called an *arithmetic series* $a_1 + a_2 + \cdots + a_n$, can be computed as follows:

$$\sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n = \frac{(a_1 + a_n) \cdot n}{2}.$$

This can be shown by noting that any two terms which have the same distance from both ends have the same sum $a_k + a_{n-k+1} = \big(a_1 + d(k-1)\big) + \big(a_1 + d(n-k)\big) = 2a_1 + d(n-1) = a_1 + a_n$, and hence

$$\begin{aligned} 2(a_1 + a_2 + \cdots + a_n) = \quad & (a_1 + a_2 + \cdots \quad + a_n) \\ & + (a_n + a_{n-1} + \cdots \quad + a_1) \\ & = (a_1 + a_n) \cdot n. \end{aligned}$$

Because the indices of an arithmetic progression may cover a different range, for example $a_3, a_4, \ldots, a_9$, the formula for the sum is best remembered as *the sum of the first and last term, times the total number of terms, divided by two.* (Or, equivalently, *the arithmetic mean (or average) of the first and last term, times the total number of terms.*) More generally, an arithmetic progression may have infinite terms. In fact, any finite arithmetic progression with at least two terms can be extended, on either side, to form an infinite arithmetic progression, in a unique way.

Geometric progressions are analogous to arithmetic progressions, except that sums and differences are replaced by products and quotients (ratios). Hence a finite sequence $a_1, a_2, \ldots, a_n$, made of nonzero numbers, is called a *geometric progression* if the ratio $r = a_{k+1}/a_k$ between each two consecutive terms is constant. Note that the ratio r is nonzero, but may possibly be negative, in which case the terms of the progression have alternating signs. Similarly as for arithmetic progressions, a sequence is a geometric progression precisely when $a_{k+1} \cdot a_{k-1} = a_k^2$ for all k which make sense, that is to say, when each term is the geometric mean of the preceding and the following term.

For a geometric progression we have

$$a_n = a_1 \cdot r^{n-1}.$$

[4] A sequence satisfying $a_{k+1} + a_{k-1} \geq 2a_k$ for all k is usually called *convex,* and *strictly convex* if $a_{k+1} + a_{k-1} > 2a_k$ for all k. Similarly, a sequence satisfying $a_{k+1} + a_{k-1} \leq 2a_k$ for all k is usually called *concave.* You may relate this to Calculus notions if you note that $a_{k+1} - 2a_k + a_{k-1}$ is a discrete analogue of the 'second derivative' of a function (and here a_k is a function of the discrete variable k).

In a finite geometric progressions, any two terms which have the same distance from both ends have the same product. Hence, arguing in a similar way as for the sum of an arithmetic progression we see that the product of all terms of a geometric progression, say made of positive real numbers for simplicity, is given by

$$\prod_{k=1}^{n} a_k = a_1 \cdot a_2 \cdots a_n = \sqrt{(a_1 a_n)^n}.$$

This can also be thought of as *the n-th power* of the *geometric mean* $\sqrt{a_1 a_n}$ of the first and last terms.

One may also consider the sum of a geometric progression, also called a *geometric series*. If $a_1, a_2, \ldots, a_n$ is a geometric progression with (common) ratio r, then

$$a_1 + a_2 + a_3 + \cdots + a_n = a_1(1 + r + r^2 + \cdots + r^{n-1}) = a_1 \frac{r^n - 1}{r - 1} = a_1 \frac{1 - r^n}{1 - r}.$$

This can also be used to compute the sum of an infinite geometric progression $a_1, a_2, \ldots$ (continuing indefinitely to the right). The corresponding geometric series converges (see the Calculus module for the meaning of this) exactly when r^n tends to zero as n tends to $+\infty$, which occurs exactly when $|r| < 1$. In that case the sum of the series is given by

$$a_1 + a_2 + a_3 + \cdots = a_1(1 + r + r^2 + r^3 + \cdots) = \frac{a_1}{1 - r}.$$

Remark (Optional: Arithmetic, geometric, and harmonic mean). We mentioned above the *arithmetic mean* $(a+b)/2$ of two numbers, and the *geometric mean* $\sqrt{ab}$ of two *positive real* numbers. Note that the geometric mean never exceeds the arithmetic mean, that is, $\sqrt{ab} \leq (a + b)/2$ for all positive real numbers a, b. In fact, because both sides are positive this inequality is equivalent to $ab \leq (a + b)^2/4$. In turn, this is equivalent to $0 \leq (a+b)^2 - 4ab$, that is, $0 \leq (a-b)^2$, which is certainly true for all positive real numbers a, b. A third type of mean occurs in some applications, the *harmonic mean* $\frac{1}{\frac{1}{2}(\frac{1}{a}+\frac{1}{b})} = \frac{2ab}{a+b}$. The harmonic mean of two positive real numbers a, b never exceeds their geometric mean (similar proof), and so we have

$$\frac{2ab}{a + b} \leq \sqrt{ab} \leq \frac{a + b}{2},$$

that is, [harmonic mean]≤[geometric mean]≤[arithmetic mean]. Note that the product of the arithmetic mean and the harmonic mean equals the square of the geometric mean; this is another (but equivalent) explanation of why the geometric mean takes an intermediate value between the other two means.

9. Periodic numbers (in decimal notation or any other base)

Example. Converting a real number with periodic decimal expansion into a fraction:

$$0.171717\cdots = 0.\overline{17} = 0.17 \cdot 1.\overline{01} = 0.17 \cdot \big(1 + (0.01) + (0.01)^2 + \cdots\big) = \frac{0.17}{1 - 0.01} = \frac{17}{99}.$$

It is easy to discover how to extend this to the most general situation of a *periodic decimal expansion* (also called a *repeated* or *recurring decimal*, for which various notations are in use) with both an *integer part* and a *pre-period*:

$$1234.56789789789\cdots = 1234.56\overline{789} = 1234.56\dot{7}8\dot{9} = \frac{123456789 - 123456}{99900}.$$

The numerator of the fraction equals

$$\text{[integer part|pre-period|period]} \quad \text{minus} \quad \text{[integer part|pre-period]},$$

ignoring the decimal dot; the denominator has as many 9s as the number of digits of the period, followed by as many 0s as the digits of the pre-period.

The procedure explained in the example shows that a real number whose decimal expansion is periodic, is actually a rational number (a fraction of integers). Obviously, a real number whose decimal expansion is finite is also a rational number. (This may actually be viewed as a special case of a periodic expansion where the period is $\dot{0}$, and the procedure for converting it to a fraction still works...) More is true: *a real number has finite or periodic decimal expansion if and only if it is rational.* We have just seen the 'only if' implication, that is, the '$\Rightarrow$' implication. To prove the opposite implication, note that when computing the decimal expansion of a rational number m/n (hence with m and n integers), that is, when performing the ordinary school division algorithm (similar to long division for polynomials), at each step at most n remainders r are possible (as $0 \leq r < n$). Once we have finished 'carrying down' all the digits from m (so the following ones would all be zeroes, which we usually do not write), sooner or later one of the remainders will have to repeat, and from that point on a whole bunch of steps of the division algorithm will have to repeat periodically. It is easier to see what happens by working out an example than explaining it in words, but it follows that the resulting decimal expansion must be periodic.

The same procedure would work in any base b, just replace the digits 9 used in the rule with the digit $b-1$. However, beware that those numbers you are writing as numerator and denominator of the fraction will be in base b, so you may then need to convert them to decimal in order to write the fraction in the ordinary way. For example,

$$(3.\dot{2}\dot{1})_7 = \frac{(321)_7 - (3)_7}{(66)_7} = \frac{(315)_7}{(66)_7} = \frac{159}{48}.$$

Note that some fractions of integers may have a finite expansion when written in some base, and a periodic infinite expansion when written in some other base:

$$\frac{1}{2} = 0.5 = (0.1)_2 = (0.\dot{1})_3 = (0.2)_4 = (0.\dot{2})_5 = (0.3)_6 = (0.\dot{3})_7 = \cdots$$

$$\frac{1}{3} = 0.\dot{3} = (0.\dot{0}\dot{1})_2 = (0.1)_3 = (0.\dot{1})_4 = (0.\dot{1}\dot{3})_5 = (0.2)_6 = \cdots$$

10. Polynomials

In this and the following sections we will consider polynomials with coefficients in a *field* F. Some examples of fields are $\mathbb{Q}$ (the field of rational numbers), $\mathbb{R}$ (the field of real numbers), $\mathbb{C}$ (the field of complex numbers). These fields satisfy $\mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$, but later we will discover some other interesting and unrelated fields. Most of what we do with polynomials in the following section does not really depend on which field we use (unless we state that explicitly), and so you may take F to be some field which is familiar to you, say $\mathbb{Q}$ or $\mathbb{R}$, and worry only later about the proper definition of an arbitrary field. Very roughly, a field is a set of 'numbers' which you can add, subtract, and multiply arbitrarily, and also divide except for dividing by zero, and where those operations satisfy the familiar calculation 'rules,' such as $a-(b-c)=a-b+c$, $ab=ba$, $a(b+c)=ab+ac$, etc.

Example. The set of integers $\mathbb{Z}$ is not a field, because division cannot be done arbitrarily: $2/3$ is not an integer. One can of course consider polynomials with integer coefficients (being special rational numbers), but the theorems which we will see may not be true, and the algorithms may not work, unless we accept to use rational numbers, which are a field.

Example. There are many different fields, and even some with a finite number of elements. The simplest example is the field with two elements, usually denoted by $\mathbb{F}_2$. (Later in the module we will learn about the field $\mathbb{F}_p$ of p elements, where p is any prime.) Its elements are two symbols $\bar{0}$ and $\bar{1}$, which add and multiply exactly as the integers 0 and 1 do, except that $\bar{1}+\bar{1}=\bar{0}$. Here are the full addition and multiplication tables in $\mathbb{F}_2$:

$+$	$\bar{0}$	$\bar{1}$
$\bar{0}$	$\bar{0}$	$\bar{1}$
$\bar{1}$	$\bar{1}$	$\bar{0}$

$\cdot$	$\bar{0}$	$\bar{1}$
$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$

One can verify that $\mathbb{F}_2$ satisfies the definition of a field (which we have not explicitly given but roughly amounts to saying that the usual properties of addition, subtraction, multiplication and division hold).

For the moment, we may think of a polynomial (in the single *indeterminate* x) as an expression of the form $f(x)=a_nx^n+\cdots+a_1x+a_0$. (Arranging them in the opposite order is just another convention in use.) The notation $f(x)$ is borrowed from Calculus to stress that one may think of the polynomial as a special type of function of the "variable" x, while $a_0, a_1, \ldots, a_n$ are to be thought of as "constants" (even though in some applications they may themselves be expressions depending on parameters, other than x). Strictly speaking, a function of this type should be called *a polynomial function,* rather than *a*

polynomial. Thinking of polynomials as functions is OK as long as one works with real coefficients, but is not quite the best in view of generalizations.

Thus, a polynomial with coefficients in the field F is an expression of the form $f(x) = a_n x^n + \cdots + a_1 x + a_0$ with $a_0, \ldots, a_n \in F$, for some n. If β is any element of F, we may *evaluate* $f(x)$ *on* β, *or for* $x = \beta$, and compute the value $f(\beta) = a_n \beta^n + \cdots + a_1 \beta + a_0$.

DEFINITION 19. The degree of a non-zero polynomial $f(x)$ is the largest integer n such that $a_n \neq 0$, and is denoted by $n = \deg(f)$.

The coefficient a_n in the definition is called *the leading coefficient*, and if $a_n = 1$ then $f(x)$ is said to be *monic*. We also call $a_n x^n$ *the leading term* of the polynomial, and a_0 *the constant term*. We do not assign a degree to the zero polynomial. (In our notation $f(x) = a_n x^n + \cdots + a_1 x + a_0$ we have not assumed that $a_n \neq 0$. This is actually convenient, and all that notation tells us is that $\deg(f(x)) \leq n$, or $f(x)$ might possibly be the zero polynomial.)

The sum of two polynomials is given by

$$(a_n x^n + \cdots + a_1 x + a_0) + (b_n x^n + \cdots + b_1 x + b_0) = (a_n + b_n)x^n + \cdots + (a_1 + b_1)x + (a_0 + b_0).$$

Note that the two polynomials need not have the same degree, and to make this formula simpler we have taken advantage of the possibility of adding zero coefficients in front of one of them to make both polynomials formally start with the same x^n.

The product of two polynomials can be computed by removing the parentheses using the distributive law, and then collecting like powers of x. Hence

$$\begin{aligned}(a_n x^n + \cdots + a_1 x + a_0) \cdot (b_m x^m + \cdots + b_1 x + b_0) \\ = a_n b_m x^{n+m} + (a_{n-1} b_m + a_n b_{m-1}) x^{n+m-1} + \cdots \\ \cdots + (a_0 b_2 + a_1 b_1 + a_2 b_0) x^2 + (a_0 b_1 + a_1 b_0) x + a_0 b_0.\end{aligned}$$

Those two formulas show that the degrees of a sum and of a product of polynomial satisfy

$$\deg\big(f(x) + g(x)\big) \leq \max\big(\deg(f(x)), \deg(g(x))\big),$$

and

$$\deg\big(f(x) \cdot g(x)\big) = \deg\big(f(x)\big) + \deg\big(g(x)\big),$$

provided that all degrees make sense, that is, unless one of the polynomials involved is the zero polynomial.

11. Polynomial division with remainder

THEOREM 20 (Division Algorithm for $F[x]$). *Let F be a field and let $f(x), g(x) \in F[x]$ with $g(x) \neq 0$. Then there exist unique polynomials $q(x), r(x) \in F[x]$ such that*

$$f(x) = g(x)q(x) + r(x)$$

where

$$\textit{either} \quad r(x) = 0 \quad \textit{or} \quad \deg\big(r(x)\big) < \deg\big(g(x)\big).$$

The polynomials $q(x)$ and $r(x)$ are called the *quotient* and the *remainder* of the division of $f(x)$ by $g(x)$.

Remark (The degree of the zero polynomial). An alternate convention is to assign a degree to the zero polynomial as well. To make things work properly, however, the zero polynomial should be assigned some negative number, say -1. Doing so, the condition

$$\text{either} \quad r(x) = 0 \quad \text{or} \quad \deg\big(r(x)\big) < \deg\big(g(x)\big)$$

in Theorem 20 can be rephrased, more simply, as

$$\deg\big(r(x)\big) < \deg\big(g(x)\big).$$

An even better convention is assigning the symbol $-\infty$ to the zero polynomial, which makes the properties given above on the degrees of a sum and a product remain true even if one of the polynomials involved is the zero polynomial (with some natural interpretations such as $(-\infty) + 3 = -\infty$, or $(-\infty) + (-\infty) = -\infty$).

We omit the proof of Theorem 20, which is just a bit tedious to write down formally. However, a proof of the *existence* of $q(x)$ and $r(x)$ is just a formal transcription of what the actual algorithm does, as illustrated in the following example.

Example. In $\mathbb{Q}[x]$, divide $f(x) = 2x^4 + x^2 - x + 1$ by $g(x) = 2x - 1$ with remainder.

$$\begin{array}{rl|llllllll}
 & & x^3 & + \tfrac{1}{2}x^2 & + \tfrac{3}{4}x & - \tfrac{1}{8} & & & \\
\hline
2x-1 & | \; 2x^4 & + 0x^3 & + x^2 & - x & + 1 & & & \\
 & 2x^4 & - x^3 & & & & & & \\
\cline{2-3}
 & & x^3 & + x^2 & & & & & \\
 & & x^3 & - \tfrac{1}{2}x^2 & & & & & \\
\cline{3-4}
 & & & \tfrac{3}{2}x^2 & - x & & & & \\
 & & & \tfrac{3}{2}x^2 & - \tfrac{3}{4}x & & & & \\
\cline{4-5}
 & & & & - \tfrac{1}{4}x & + 1 & & & \\
 & & & & - \tfrac{1}{4}x & + \tfrac{1}{8} & & & \\
\cline{5-6}
 & & & & & \tfrac{7}{8} & & &
\end{array}$$

Therefore, we have $q(x) = x^3 + \frac{1}{2}x^2 + \frac{3}{4}x - \frac{1}{8}$ and $r(x) = \frac{7}{8}$. Another perfectly acceptable (and perhaps even preferable) answer is

$$2x^4 + x^2 - x + 1 = (2x - 1) \cdot \left(x^3 + \frac{1}{2}x^2 + \frac{3}{4}x - \frac{1}{8}\right) + \frac{7}{8}.$$

Note that the algorithm stops as soon as we obtain a *remainder* which is zero or has degree less than the degree of $g(x)$, as stated in Theorem 20. For example, if we stopped two steps too early (omitting the last four lines) we would obtain that $f(x) = g(x)q_1(x) + r_1(x)$ with $q_1(x) = x^3 + \frac{1}{2}x^2$ and $r_1(x) = \frac{3}{2}x^2 - x + 1$, which is a true equality, but those are not the correct quotient and remainder because $\deg\big(r_1(x)\big) = 2$ is not less than $\deg(2x - 1) = 1$, and $r_1(x)$ is not zero either.

Note that, although the two original polynomials in the above example had integer coefficients, we had to use rational numbers in the course of the calculation, and also to express the final result. That example shows that division with remainder of polynomials in $\mathbb{Z}[x]$ would simply not work, and the reason is that $\mathbb{Z}$ is not a field (as we cannot divide arbitrarily by non-zero integers). Division of polynomials with integer coefficients does work in a restricted situation, namely when the polynomial we are dividing by is monic (that is, it has leading coefficient 1), as in the next example.

Example. In $\mathbb{Q}[x]$, divide $f(x) = 2x^4 - x^3 + 3x^2 + x - 2$ by $g(x) = x^2 - 2x + 2$ with remainder.

$$
\begin{array}{rl}
 & 2x^2 + 3x + 5 \\
\hline
x^2 - 2x + 2 \;\big| & 2x^4 - x^3 + 3x^2 + x - 2 \\
 & \underline{2x^4 - 4x^3 + 4x^2} \\
 & \quad 3x^3 - x^2 + x \\
 & \quad \underline{3x^3 - 6x^2 + 6x} \\
 & \qquad\quad 5x^2 - 5x - 2 \\
 & \qquad\quad \underline{5x^2 - 10x + 10} \\
 & \qquad\qquad\quad 5x - 12
\end{array}
$$

Therefore $q(x) = 2x^2 + 3x + 5$ and $r(x) = 5x - 12$. Or, more explicitly,

$$2x^4 + -x^3 + 3x^2 + x - 2 = (x^2 - 2x + 2) \cdot (2x^2 + 3x + 5) + (5x - 12).$$

A proof of the *uniqueness* of $q(x)$ and $r(x)$ in Theorem 20 is easier to write down formally than a proof of their existence.

Proof of uniqueness of $q(x)$ and $r(x)$ in Theorem 20. Suppose that the division can be done in two ways,

$$f(x) = g(x)q(x) + r(x) \qquad \text{and} \qquad f(x) = g(x)q_1(x) + r_1(x),$$

with both $r(x)$ and $r_1(x)$ satisfying the required condition. Then we claim that $q_1(x) = q(x)$ and $r_1(x) = r(x)$. In fact, putting together the two equalities we obtain

$$g(x)q(x) + r(x) = f(x) = g(x)q_1(x) + r_1(x),$$

and hence

$$g(x)[q_1(x) - q(x)] = r(x) - r_1(x).$$

If the right-hand side were different from zero, then its degree would be less than the degree of $g(x)$. However, the left-hand side, if nonzero, would have degree $\deg(g(x)) + \deg[q_1(x) - q(x)] \geq \deg(g(x))$. This is impossible, and so we have to conclude that each side of the equality is zero. This implies that $r_1(x) = r(x)$, and $q_1(x) = q(x)$ (because $g(x)$ is not the zero polynomial), as we wanted to prove. □

Remark 21. Although division with remainder for polynomials is conceptually very similar to division with remainder for integers, with polynomials we have no analogue of the variant of division with $-b/2 < r \leq b/2$: there is only one way to do division with remainder for polynomials.

Lecture notes of Algebra. Week 4

12. The Remainder Theorem and the Factor Theorem

Let $f(x) \in F[x]$, and $\alpha \in F$. We say that α is a *root* (or a *zero*) of $f(x)$ if $f(\alpha) = 0$. In other words, if we obtain zero after substituting α for x in $f(x) = a_n x^n + \cdots + a_1 x + a_0$, that is, computing $f(\alpha) = a_n \alpha^n + \cdots + a_1 \alpha + a_0$.

We say that a polynomial $g(x)$ divides $f(x)$ if there is another polynomial $h(x)$ such that $f(x) = g(x) \cdot h(x)$. This occurs exactly when dividing $f(x)$ by $g(x)$ we obtain zero as the remainder.

Lemma 22 (The Remainder Theorem and the Factor Theorem). *Let F be a field, $0 \neq f(x) \in F[x]$, $\alpha \in F$. Then*

(1) *$f(\alpha)$ equals the remainder of the division of $f(x)$ by $x - \alpha$;*
(2) *α is a root of $f(x)$ if, and only if, $x - \alpha$ divides $f(x)$.*

Proof. Dividing $f(x)$ by $(x - \alpha)$ we obtain $f(x) = (x - \alpha) \cdot q(x) + r$, where r is either zero or a polynomial of degree less than 1, hence a constant $r \in F$ in both cases. Evaluating on α we find $f(\alpha) = (\alpha - \alpha) \cdot q(\alpha) + r = r$, which proves the remainder theorem.

The Factor Theorem is an immediate consequence: the equality $f(x) = (x-\alpha) \cdot q(x) + r$ shows that $x - \alpha$ divides $f(x)$ (exactly) if, and only if, $r = 0$; but $r = f(\alpha)$ according to the remainder theorem. □

A nice application of the Factor Theorem arises when $f(x) = x^n \pm a^n$, where $a \neq 0$ is a constant. Because $f(a) = a^n \pm a^n$ and $f(-a) = (-1)^n a^n \pm a^n$ the Factor Theorem implies:

- $x^n - a^n$ is always divisible by $x - a$;
- $x^n - a^n$ is divisible by $x + a$ exactly when n is even;
- $x^n + a^n$ is divisible by $x + a$ exactly when n is odd;
- $x^n + a^n$ is never divisible by $x - a$.

For example,

$$x^2 - a^2 = (x - a)(x + a)$$
$$x^3 - a^3 = (x - a)(x^2 + ax + a^2)$$
$$x^3 + a^3 = (x + a)(x^2 - ax + a^2)$$
$$x^4 - a^4 = (x - a)(x^3 + ax^2 + a^2 x + a^3) = (x + a)(x^3 - ax^2 + a^2 x - a^3)$$

Note, however, that if we had to factorise $x^4 - a^4$ it would be smarter to think of it as $(x^2)^2 - (a^2)^2$, and so

$$x^4 - a^4 = (x^2 - a^2)(x^2 + a^2) = (x - a)(x + a)(x^2 + a^2).$$

Note also that, if $a \in \mathbb{R}$ and we work over the real numbers, $x^2 + a^2$ cannot be further factorised, again because of the Factor Theorem, since it cannot have any real roots: whatever real value we assign to x will make $x^2 + a^2$ a positive number, hence never zero. Over the complex numbers, however, we have

$$x^4 - a^4 = (x - a)(x + a)(x - ai)(x + ai),$$

as ai and $-ai$ are the square roots of $-a^2$. (Every polynomial in $\mathbb{C}[x]$ can be factorised into a product of factors of degree 1, this is called *the Fundamental Theorem of Algebra*.)

Similarly, if we have to factorise $x^6 - a^6$, we best proceed as follows:

$$\begin{aligned} x^6 - a^6 &= (x^3)^2 - (a^3)^2 \\ &= (x^3 - a^3)(x^3 + a^3) \\ &= (x - a)(x^2 + ax + a^2)(x + a)(x^2 - ax + a^2). \end{aligned}$$

Over the real numbers the two quadratic factors cannot be further factorised, because they have negative discriminant $a^2 - 4a^2 = -3a^2$. Starting, instead, with

$$\begin{aligned} x^6 - a^6 &= (x^2)^3 - (a^2)^3 \\ &= (x^2 - a^2)(x^4 + a^2x^2 + a^4) \\ &= (x - a)(x + a)(x^4 + a^2x^2 + a^4), \end{aligned}$$

would not have been as good, because none of the above rules applies directly to further factorise the factor of degree four. However, there is another trick which can be very useful in certain situations, namely,

$$\begin{aligned} x^4 + a^2x^2 + a^4 &= (x^4 + 2a^2x^2 + a^4) - a^2x^2 \\ &= (x^2 + a^2)^2 - (ax)^2 \\ &= (x^2 - ax + a^2)(x^2 + ax + a^2), \end{aligned}$$

and so we would recover the complete (or full) factorisation of $x^6 - a^6$ as before. This trick also allows us to factorise $x^6 + a^6$, where writing it as $(x^3)^2 + (a^3)^2$ would have been a dead end:

$$x^6 + a^6 = (x^2 + a^2)(x^4 - a^2x^2 + a^4).$$

One could show that this is a complete factorisation over the rational numbers (if a is rational), but over the real numbers the same trick factorises the factor of degree four:

$$\begin{aligned} x^4 - a^2x^2 + a^4 &= (x^4 + 2a^2x^2 + a^4) - 3a^2x^2 \\ &= (x^2 + a^2)^2 - (\sqrt{3}ax)^2 \\ &= (x^2 - a\sqrt{3}x + a^2)(x^2 + a\sqrt{3}x + a^2). \end{aligned}$$

13. Ruffini's rule

Because of the Factor Theorem, the very special case of polynomial division where we divide by a binomial of the form $x - a$, for some constant a, is important. In this case the ordinary division algorithm is very sparse, and all the numbers involved can be arranged in a more compact notation, which we illustrate by an example: to divide $f(x) = x^4 + 3x^3 - 5x - 10$ by $x - 2$ we write

$$\begin{array}{c|cccc|c} & 1 & 3 & 0 & -5 & -10 \\ 2 & & 2 & 10 & 20 & 30 \\ \hline & 1 & 5 & 10 & 15 & 20 \end{array}$$

and conclude that $x^4 + 3x^3 - 5x - 10 = (x^3 + 5x^2 + 10x + 15)(x - 2) + 20$. This method of performing the division is called *Ruffini's method* (or *Ruffini's rule*). (An extension of it exists, called *synthetic division,* which allows division by any monic polynomial rather than a special binomial $x - a$.)

According to the Factor Theorem, the remainder 20 of the division equals $f(2)$, and so this algorithm can also be used to *evaluate* a polynomial $f(x)$ on a number a (that is, to compute $f(a)$). This algorithm (called *Horner scheme,* but equivalent to Ruffini's rule), which really amounts to rewriting $x^4 + 3x^3 - 5x - 10$ as

$$((((x + 3)x + 0)x - 5)x - 10,$$

is more efficient than the obvious one (computing the various powers of 2, multiplying them by the corresponding coefficients, and then adding up the results), as it requires the same number of addition, but about half as many multiplications. Not a drastic saving, but significant.

Exercise. For a generic polynomial $f(x)$ of degree n (that is, avoiding special cases where some coefficient is zero), exactly how many additions and how many multiplications are required to compute $f(a)$ by the two methods?

Example. Continuing with a previous example, we already know that $f(x) = x^n - a^n$ is divisible by $x - a$. In fact, dividing by means of Ruffini's rule we find remainder zero:

$$\begin{array}{c|ccccc|c} & 1 & 0 & 0 & \cdots & 0 & -a^n \\ a & & a & a^2 & \cdots & a^{n-1} & a^n \\ \hline & 1 & a & a^2 & \cdots & a^{n-1} & 0 \end{array}$$

But Ruffini's rule also gives us the quotient, and so we find

$$x^n - a^n = (x - a)(x^{n-1} + ax^{n-2} + \cdots + a^{n-2}x + a^{n-1}).$$

Of course this identity could be easily verified directly by executing the multiplication on the RHS, but the point is that Ruffini's rule produces the quotient very quickly. When

n is odd (and only then) a similar identity

$$x^n + a^n = (x + a)(x^{n-1} - ax^{n-2} + \cdots - a^{n-2}x + a^{n-1})$$

can be obtained by applying Ruffini's rule to divide $f(x) = x^n + a^n$ by $x + a$, but it is quicker to deduce the identity by replacing a with $-a$ in the previous identity.

14. Expansion of a polynomial in terms of $x - a$

The following example illustrates how iterating Ruffini's rule we can expand a polynomial in x into powers of $x - a$, that is, if we like, into *a polynomial in* $x - a$. Say we want to expand $f(x) = x^3 + 2x^2 - x - 3$ into powers of $x - 2$. Then we do the following: we divide $f(x)$ by $x - 2$, then we divide the resulting quotient by $x - 2$, then we divide the resulting quotient by $x - 2$, and so on until the quotient is zero.

	1	2	−1	−3
2		2	8	14
	1	4	7	**11**
2		2	12	
	1	6	**19**	
2		2		
	1	**8**		
2				
	1			

The final result of this calculation is that

$$x^3 + 2x^2 - x - 3 = 1(x - 2)^3 + 8(x - 2)^2 + 19(x - 2) + 11.$$

The reason why it works is that the term 11 can be obtained as the remainder of dividing the polynomial by $x - 2$, which is done via Ruffini's rule. The quotient $x^2 + 4x + 7$ of this division is eventually going to be written as $1(x - 2)^2 + 8(x - 2) + 19$, and hence 19 can be obtained as the quotient of dividing it by $x - 2$. And so on.

Of course an alternative way of expanding a polynomial into powers of $x - a$ is substituting $x = y + a$ into it, then expanding the various powers of $y + a$ involved, thus converting it into a polynomial in y after the appropriate simplifications, and finally set $y = x - a$. This procedure, however, involves more operations and hence is computationally less efficient.[5]

[5]The way described of expanding a polynomial in terms of $x - a$ is analogous to the efficient way to convert an integer from decimal to another base b, which was described earlier in the notes.

15. Divisibility, GCD, and the Euclidean algorithm for polynomials

Divisibility, GCD and lcm, and the Euclidean algorithm, work for polynomials with coefficients in a field F very much the same as they do for integers, with only small adjustments. To begin with, divisibility, divisors, etc., are defined in the same way:

Definition 23 (Divisibility for polynomials). Let $f(x)$ and $g(x)$ be polynomials with coefficients in a field F. We say that $g(x)$ divides $f(x)$, and we write $g(x) \mid f(x)$, if there is a polynomial $h(x) \in F[x]$ such that $f(x) = g(x) \cdot h(x)$.

With polynomials, if $f(x) \mid g(x)$ and $g(x) \mid f(x)$, then $g(x) = c \cdot f(x)$ for some nonzero constant c. In fact, the nonzero constants play the same role in $F[x]$ as ± 1 play in $\mathbb{Z}$: they are exactly the elements which are *invertible,* that is, which have an *inverse* (belonging to the same set). This means that the only polynomials $c(x)$ such that there is a polynomial $d(x)$ with $c(x) \cdot d(x) = 1$, are exactly the nonzero constant polynomials $c(x)$ (which, being constants, we may simply write as c). To show this rigorously, taking the degrees in the equality $c(x) \cdot d(x) = 1$ gives us $\deg\big(c(x)\big) + \deg\big(d(x)\big) = 0$, which can only happen if $\deg\big(c(x)\big) = \deg\big(d(x)\big) = 0$, that is, $c(x)$ is a nonzero constant polynomial (and so is $d(x)$).

The greatest common divisor of two polynomials $f(x)$ and $g(x)$ is defined in the same way as for the integers:

Definition 24 (Greatest common divisor). Let $f(x)$ and $g(x)$ be polynomials with coefficients in a field F (hence $f(x), g(x) \in F[x]$, in a formula). A polynomial $d(x) \in F[x]$ is called *a greatest common divisor* of $f(x)$ and $g(x)$ if

(1) $d(x)$ divides $f(x)$ and $g(x)$, and
(2) if $c(x) \in F[x]$ is any polynomial which divides both $f(x)$ and $g(x)$, then $c(x)$ divides $d(x)$.

Remark. Beware that this is different from the definition given in Part II, Chapter 2, of the recommended book by Childs. The definition given there is more in the style of a 'school' definition, but the one we give here has the advantage of being (essentially) the same for integers, for polynomials, and for other contexts: whenever you have a concept of divisibility, there is a concept of *a* greatest common divisor (which need not be precisely unique).

A greatest common divisor of $f(x)$ and $g(x)$, denoted by $\big(f(x), g(x)\big)$, is only unique up to multiplying it by a nonzero constant. Hence saying that the GCD of two polynomials is $x+3$ is equivalent to saying that it is $2x+6$, or $\frac{1}{3}x+1$, etc. Among all those equivalent GCD's one usually chooses the one which is monic. (This is similar to choosing the positive greatest common divisors of two integers, rather than its opposite.) As we do

for integers, if two polynomials $f(x)$ and $g(x)$ have greatest common divisor 1 then we say that they are *coprime*.

The Euclidean algorithm and the extended Euclidean algorithm work for polynomials in the same way as for the integers. Hence, given two polynomials $f(x)$ and $g(x)$, with $\deg\big(f(x)\big) \geq \deg\big(g(x)\big)$ (otherwise we just swap the two polynomials), we start the algorithm by dividing the first polynomials by the second:

$$f(x) = g(x)q_1(x) + r_1(x), \qquad \text{with } \deg\big(r_1(x)\big) < \deg\big(g(x)\big).$$

Then we divide $g(x)$ by the first remainder $r_1(x)$,

$$g(x) = r_1(x)q_2(x) + r_2(x), \qquad \text{with } \deg\big(r_2(x)\big) < \deg\big(r_1(x)\big),$$

and repeat the procedure until some remainder is zero. The last nonzero remainder is then the GCD of the two polynomials. As in the case of integers, this is justified by noting the following basic fact: if $f(x) = g(x)q(x)+r(x)$ then $\mathrm{GCD}\big(f(x), g(x)\big) = \mathrm{GCD}\big(g(x), r(x)\big)$. Hence if $r_i(x)$ is the last nonzero remainder, then $\mathrm{GCD}\big(f(x), g(x)\big) = \mathrm{GCD}\big(g(x), r_1(x)\big) = \mathrm{GCD}\big(r_1(x), r_2(x)\big) = \cdots = \mathrm{GCD}\big(r_i(x), 0\big) = r_i(x)$.

The number of divisions required by the Euclidean algorithm on polynomials is at most the lower of the degrees of the two polynomials. This is easy to see as the degree of each remainder is less than the degree of the previous remainder.

Example. We compute the GCD of the polynomials $x^3 + 2x^2 + x$ and $x^2 + x - 1$ using the Euclidean algorithm:

$$\begin{aligned} x^3 + 2x^2 + x &= (x^2 + x - 1) \cdot (x + 1) + (x + 1) \\ x^2 + x - 1 &= (x + 1) \cdot x - 1. \end{aligned}$$

The remainder of the second division is -1, so there is no point in doing a third division, as dividing by -1 (or by 1, or $2/3$, or any nonzero rational number) would give remainder zero. Hence the last nonzero remainder is -1, and so GCD of x^3+2x^2+x and x^2+x-1 is 1. In words, those polynomials are coprime.

Example. We compute the GCD of the polynomials $x^{3n} - 1$ and $x^{2n} - 1$, where n is any positive integer. The Euclidean algorithm reads:

$$\begin{aligned} x^{3n} - 1 &= (x^{2n} - 1) \cdot x^n + (x^n - 1) \\ x^{2n} - 1 &= (x^n - 1) \cdot (x^n + 1). \end{aligned}$$

Hence the last nonzero remainder is $x^n - 1$, and so GCD of $x^{3n} - 1$ and $x^{2n} - 1$ is $x^n - 1$.

We should have known from the start that $x^n - 1$ divides both polynomials. In fact $x^{3n} - 1 = (x^n - 1)(x^{2n} + x^n + 1)$ follow from the general identity $x^3 - a^n = (x - a)(x^2 + ax+a^2)$. Similarly, we should have known that $x^{2n}-1 = (x^n-1)(x^n+1)$. Knowing these factorisations, another way to prove that $x^2 - 1$ is actually the *greatest* common divisor

of $x^{3n}-1$ and $x^{2n}-1$ (rather than just *a* common divisor) would then be showing that $x^{2n}+x^n+1$ and x^n+1 are coprime. Of course we can do that by applying the Euclidean algorithm, which in this case consists of a single division: $x^{2n}+x^n+1=(x^n+1)\cdot x^n+1$. This tells us that their GCD is 1, and so the GCD of $x^{3n}-1$ and $x^{2n}-1$ is x^n-1.

REMARK. One can actually prove that for any positive integers m and n the greatest common divisor of x^m-1 and x^n-1 is $x^{(m,n)}-1$ (where the exponent (m,n) means the GCD of m and n, as usual).

The conclusion of the extended Euclidean algorithm can be formally stated as Bézout's Lemma for polynomials:

LEMMA 25 (Bézout's Lemma for polynomials). *Let $f(x),g(x)\in F[x]$, where F is a field, and let $d(x)=\big(f(x),g(x)\big)$ be their greatest common divisor. Then there exist polynomials $u(x),v(x)\in F[x]$ such that*

$$f(x)\,u(x)+g(x)\,v(x)=d(x).$$

It is not difficult to show that if neither $f(x)$ or $g(x)$ is the zero polynomial then the polynomials $u(x)$ and $v(x)$ produced by the extended Euclidean algorithm satisfy

$$\deg\big(u(x)\big)<\deg\big(g(x)\big)\qquad\text{and}\qquad\deg\big(v(x)\big)<\deg\big(f(x)\big).$$

EXAMPLE. Reading the divisions in the previous example backwards we find:

$$\begin{aligned}1&=-(x^2+x-1)+(x+1)\cdot x\\&=-(x^2+x-1)+[(x^3+2x^2+x)-(x^2+x-1)\cdot(x+1)]\cdot x\\&=(x^3+2x^2+x)\cdot x+(x^2+x-1)\cdot[-1-(x+1)x]\\&=(x^3+2x^2+x)\cdot x+(x^2+x-1)\cdot(-x^2-x-1).\end{aligned}$$

So we have found find two polynomials $u(x)$ and $v(x)$ such that

$$(x^3+2x^2+x)\cdot u(x)+(x^2+x-1)\cdot v(x)=1,$$

namely,

$$u(x)=x,\quad\text{and}\quad v(x)=-x^2-x-1.$$

Note that these polynomials satisfy

$$\frac{1}{(x^3+2x^2+x)(x^2+x-1)}=\frac{u(x)}{x^2+x-1}+\frac{v(x)}{x^3+2x^2+x}.$$

Hence the extended Euclidean algorithm has allowed us to write the fraction on the LHS as a sum of the two 'simpler' fractions on the RHS. Note that in each of the two fractions on the RHS the numerator has degree strictly less than the denominator. This is a particular instance of the general fact on the degrees of $u(x)$ and $v(x)$ mentioned before the example.

EXAMPLE. We compute the *monic* greatest common divisor $d(x)$ of $x^3 - x^2 + x - 6$ and $x^3 + x - 10$:

$$\begin{aligned} x^3 - x^2 + x - 6 &= (x^3 + x - 10) \cdot 1 + (-x^2 + 4) \\ x^3 + x - 10 &= (x^2 - 4) \cdot x + (5x - 10) \\ x^2 - 4 &= (x - 2)(x + 2). \end{aligned}$$

The last nonzero remainder is $5x - 10 = 5(x - 2)$, and so the *monic* GCD is $d(x) = (x^3 - x^2 + x - 6, x^3 + x - 10) = x - 2$. Of course it would also be correct to say that a GCD is $5x - 10$, as much as $\frac{1}{2}x - 2$, or $-\frac{2}{3}x + \frac{4}{3}$, etc., but as a standard way of choosing one we have asked for the *monic* GCD, which is $x - 2$.

Now we carry out the extended part of the Euclidean algorithm, by reading those divisions backwards, and we find

$$\begin{aligned} 5x - 10 &= (x^3 + x - 10) - (x^2 - 4) \cdot x \\ &= (x^3 + x - 10) + [(x^3 - x^2 + x - 6) - (x^3 + x - 10) \cdot 1] \cdot x \\ &= (x^3 + x - 10) \cdot x - (x^3 + x - 10) \cdot (x - 1) \end{aligned}$$

So we have found polynomials $u(x)$ and $v(x)$ whose existence is stated in Bézout's Lemma: $u(x) = \frac{1}{5}x$ and $v(x) = -\frac{1}{5}(x - 1)$ satisfy

$$(x^3 - x^2 + x - 6) \cdot u(x) + (x^3 + x - 10) \cdot v(x) = d(x) = x - 2.$$

EXAMPLE. When the GCD of two polynomials is not 1, as in the previous example, there is an alternative way to proceed with the extended Euclidean algorithm in order to have simpler calculations: once we have found that $(x^3 - x^2 + x - 6, x^3 + x - 10) = x - 2$, we may divide both polynomials by their GCD $x - 2$ and factorise them as follows:

$$\begin{aligned} x^3 - x^2 + x - 6 &= (x - 2)(x^2 + x + 3) \\ x^3 + x - 10 &= (x - 2)(x^2 + 2x + 5). \end{aligned}$$

We see from these factorisations that there was nothing special about the 5 coefficients in the GCD $5x - 10$ which we found using the Euclidean algorithm on the original polynomial, as there is no trace of that in the factorisations.

Now we carry out the extended Euclidean algorithm using the quotients by $x - 2$ instead of our original polynomials:

$$\begin{aligned} x^2 + x + 3 &= (x^2 + 2x + 5) \cdot 1 + (-x - 2) \\ x^2 + 2x + 5 &= (x + 2) \cdot x + 5. \end{aligned}$$

Hence the GCD of those polynomials is 1 (or 5, or 2/3 if you like, they all mean the same in this polynomial context, because all nonzero constants are invertible in $\mathbb{Q}[x]$). Reading

the divisions backwards we find:

$$\begin{aligned} 5 &= (x^2+2x+5) - (x+2)\cdot x \\ &= (x^2+2x+5) + [(x^2+x+3) - (x^2+2x+5)\cdot 1]\cdot x \\ &= (x^2+x+3)\cdot x - (x^2+2x+5)\cdot(x-1). \end{aligned}$$

So we have found the same $u(x) = \frac{1}{5}x$ and $v(x) = -\frac{1}{5}(x-1)$ as before. In fact, those polynomials satisfy

$$(x^2+x+3)\cdot u(x) + (x^2+2x+5)\cdot v(x) = 1,$$

and if we multiply both sides of this equality by $x-2$ we recover

$$(x^3-x^2+x-6)\cdot u(x) + (x^3+x-10)\cdot v(x) = x-2,$$

which we obtained the first time.

Note that *it would not be possible* to find polynomials $s(x)$ and $t(x)$ such that

$$(x^3-x^2+x-6)\cdot s(x) + (x^3+x-10)\cdot t(x) = 1,$$

because x^3-x^2+x-6 and x^3+x-10 are not coprime. In fact, each of them is a multiple of $x-2$, so the left-hand side is as well, but the right-hand side 1 is not, so that is impossible.

Lecture notes of Algebra. Week 5

16. Irreducible polynomials, and unique factorisation

The four Arithmetical Lemmas for integers remain true for polynomials, after making the obvious changes in terminology, and can be proved in the same way using Bézout's Lemma. In particular, the most important of them, Arithmetical Lemma B, is as follows: *if the polynomials $f(x)$ and $g(x)$ are coprime, and $f(x)$ divides the product $g(x) \cdot h(x)$, then $f(x)$ divides $h(x)$.*

Prime polynomials are usually rather called *irreducible* polynomials. As was the case for the integers, one has to exclude the zero polynomial and the invertible polynomials (which are the analogues of ± 1 in the integers) from the definition of irreducible. Because the invertible polynomials are the nonzero constants (the polynomials of degree zero), the definition of irreducible will only apply to non-constant polynomials, that is, to polynomials of positive degree.

Definition 26. A non-constant polynomial $f(x) \in F[x]$ is *reducible* in $F[x]$ if it can be written as $f(x) = g(x)\,h(x)$, where $g(x)$ and $h(x)$ are polynomials in $F[x]$ of positive degree (or, equivalently, $\deg(g(x))$ and $\deg(h(x))$ are smaller than $\deg(f(x))$; or, equivalently again, where $0 < \deg(g(x)) < \deg(f(x))$); it is *irreducible* in $F[x]$ if it is not reducible.

Because $\deg\big(g(x)\,h(x)\big) = \deg\big(g(x)\big)$, any polynomial of degree 1, hence of the form $ax + b$ with $a \neq 0$, is always irreducible, as 1 cannot be written as a sum of two positive integers.

Note that the notions of reducible and irreducible depend on the field in which we view the coefficients of our polynomial: $x^2 + 1$ is irreducible as a polynomial in $\mathbb{R}[x]$, but not as a polynomial in $\mathbb{C}[x]$, because $x^2 + 1 = (x - i)(x + i)$. To stress which field F is being used, one usually specifies *irreducible in $F[x]$*, or also *irreducible over F*. (Hence $x^2 + 1$ is irreducible over $\mathbb{R}$, but reducible over $\mathbb{C}$.)

Theorem 14 on unique factorisation in the integers has an analogue for polynomials.

Theorem 27 (Unique Factorisation Theorem for polynomials). *Every polynomial of positive degree (which is the same as saying* non-constant*) over a field F factorises into a product of irreducible polynomials (irreducible over the same field F).*

Also, the factorisation is essentially *unique, namely, unique up to permuting the factors, but also to multiplying each irreducible factor by some nonzero constant (that is, by some invertible polynomial).*

For example,

$$2x^2 + 10x + 12 = 2(x+2)(x+3) = (2x+4)(x+3) = (x+2)(2x+6)$$
$$= (3x+6)\Big(\frac{2}{3}x + 2\Big), \qquad \text{and so on.}$$

We call the factorisation into a product of irreducible polynomials the *complete factorisation* of $f(x)$ over F (or in $F[x]$). Note that, once again, with polynomials it is essential to state over which field we are working, because the answer may be different over different fields.

EXAMPLE. The polynomial $x^4 - 3x^2 + 2 \in \mathbb{Q}[x]$ can be written in three essentially different ways

$$x^4 - 3x^2 + 2 = (x^2-1)(x^2-4) = (x^2 - 3x + 2)(x^2 + 3x + 2) = (x^2 - x - 2)(x^2 + x - 2)$$

as the product of two polynomials of degree 2. However, this does not contradict the above theorem on unique factorisation because none of those quadratic factors is irreducible over $\mathbb{Q}$, in fact $x^4 - 3x^2 + 2 = (x-1)(x+1)(x-2)(x+2)$, and the above quadratic factors are obtained by pairing and multiplying together the irreducible factors in different ways.

EXAMPLE. The polynomial $x^3 - 2$ is irreducible in $\mathbb{Q}[x]$. (This will be justified in later sections.) In $\mathbb{R}[x]$ it factorises as

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

This is a complete factorisation in $\mathbb{R}[x]$ because the quadratic factor is irreducible. In fact, its discriminant is $\sqrt[3]{2}^2 - 4 \cdot \sqrt[3]{2}^2 = -3\sqrt[3]{2}^2 < 0$, and so that quadratic polynomial has no real roots. However, in $\mathbb{C}[x]$ the polynomial $x^3 - 2$ factorises as

$$x^3 - 2 = (x - \sqrt[3]{2})(x - \omega\sqrt[3]{2})(x - \bar{\omega}\sqrt[3]{2}),$$

where $\omega = (-1 \pm i\sqrt{3})/2$.

17. Quadratic polynomials

You should know well from school how to find the roots of a quadratic polynomial $ax^2 + bx + c$ (hence with $a \neq 0$, otherwise it would not be quadratic), which means the same as finding the solutions of the corresponding equation $ax^2 + bx + c = 0$. In fact, the trick of *completing the square* $ax^2 + bx$ at the left-hand side brings the equation to the equivalent form [6]

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

[6] To be precise, this works over any field $\mathbb{F}$ where $2 \neq 0$. The meaning of this unfamiliar condition will be clarified in a later algebra course, but note that it is not satisfied when $F = \mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$, the field of 2 elements, where $[2] = [1+1]$ is the same as $[0]$.

The numerator of the fraction at the right-hand side is the *discriminant* of the quadratic equation (or polynomial), and the existence of the solutions depends on whether it is a square in the field F under consideration (which is the same as being nonnegative if $F = \mathbb{R}$, but may be a different condition otherwise). If $b^2 - 4ac$ is not a square of an element of F, then no $x \in F$ can make the above equality true, and so the polynomial has no root in F. If $b^2 - 4ac$ is a square of an element of F, which means that it has a square root in F, denote it by $\sqrt{b^2 - 4ac}$, then the equation becomes

$$\left(x + \frac{b}{2a}\right)^2 - \left(\frac{\sqrt{b^2 - 4ac}}{2a}\right)^2 = 0$$

and, in turn,

$$\left(x - \frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(x - \frac{-b - \sqrt{b^2 - 4ac}}{2a}\right) = 0.$$

In conclusion, one may distinguish two cases:

- if $b^2 - 4ac$ is not the square of an element of F, and so does not have a square root in F, then the polynomial has no root in F;
- if $b^2 - 4ac$ has a square root in F, then the polynomial has roots given by the familiar formula
$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a};$$
of course the two roots coincide when $b^2 - 4ac = 0$ (which is sometimes stated separately as a third case). [7]

These cases can be expressed in terms of reducibility (over a field F containing all coefficients) of the quadratic polynomial $ax^2 + bx + c$ (with $a \neq 0$):

- it is irreducible if $b^2 - 4ac$ does not have a square root in F;
- it is reducible if $b^2 - 4ac$ has a square root in F; in fact in that case we have $ax^2 + bx + c = a(x - \alpha_1)(x - \alpha_2)$, where $\alpha_1, \alpha_2 = (-b \pm \sqrt{b^2 - 4ac})/(2a)$ are its roots (in any order); of course $\alpha_1 = \alpha_2$ when $b^2 - 4ac = 0$.

EXAMPLE. A quadratic polynomial $ax^2+bx+c \in \mathbb{R}[x]$ (hence with $a \neq 0$) is irreducible exactly when its discriminant b^2-4ac has no square roots in $\mathbb{R}$, hence exactly when b^2-4ac is negative.

EXAMPLE. Because any complex number has square roots in $\mathbb{C}$, quadratic polynomials in $\mathbb{C}[x]$ are always reducible (and hence factorise completely into products of two polynomials of degree one).

[7] Note that whenever it is convenient to collect a factor 2 from the coefficient b, for example when b is an even integer, or, say, $b = 6\sqrt{5} = 2 \cdot 3\sqrt{5}$, etc., it may be easier to use the slightly simpler formula $x = \left(-B \pm \sqrt{B^2 - ac}\right)/a$ for the roots of the polynomial $ax^2 + 2Bx + c$ (that is, where $b = 2B$).

EXAMPLE. The polynomial x^2-2 is irreducible over $\mathbb{Q}$, but reducible over $\mathbb{R}$, because $x^2-2=(x-\sqrt{2})(x+\sqrt{2})$, and $\sqrt{2}\notin\mathbb{Q}$, which means that $\sqrt{2}$ is irrational (see below, or as an application of the Rational Root Test later).

(OPTIONAL) PROOF THAT $\sqrt{2}$ IS IRRATIONAL. We need to show that $\sqrt{2}$ cannot be equal to any fraction a/b of integers. Suppose it is. (In these cases one says *suppose, for a contradiction,* and then try to show that this assumption does lead to a contradiction, meaning, proving that some fact would have to be both true and false.)

In that case the fraction can certainly be reduced to simplest terms, and so we may assume that $\sqrt{2}=a/b$, with $a,b\in\mathbb{Z}$, clearly with $b\neq 0$, and the further condition $(a,b)=1$. Multiplying by b and squaring we find $2b^2=a^2$. Hence 2 divides $a^2=a\cdot a$, but because 2 is prime it follows that 2 divides a, and hence $a=2c$ for some integer c. Substituting into our equation we find $2b^2=4c^2$, whence $b^2=2c^2$. Hence 2 divides b^2, and because 2 is prime it divides b. So we have found that 2 divides both a and b, and hence 2 divides $(a,b)=1$. But this is certainly false, and we have found the desired contradiction. (We have concluded that 2 divides 1, but at the same time we know that 2 does not divide 1.)

The only way to resolve the contradiction is to admit that we made a false assumption at the beginning, namely, in assuming that $\sqrt{2}$ is equal to some fraction a/b of integers. Hence this is not possible, which means that $\sqrt{2}$ is irrational. □

18. The maximum number of roots of a polynomial

How many roots does a polynomial have? Well, the zero polynomial has all the roots we want (any number is a root), so we look at non-zero polynomials. A polynomial of degree 1 has always exactly one root, namely, $-b/a$ if the polynomial is $ax+b$ (with $a\neq 0$ otherwise it would not have degree 1). For a polynomial of degree 2 the answer may depend on the field where we are viewing the coefficients: it may have two roots, or one root (which we may think of a double root, but we count only once if we meant to ask about how many *distinct* roots), or none. In any case, at most two. Here is a more general result, which we can prove using the Factor Theorem.

THEOREM 28. *A polynomial of degree $n\geq 0$ over a field F has at most n roots in F.*

PROOF. Let f be a polynomial of degree $n\geq 0$. If f has no roots at all in F, we are done, as $0\leq n$ (so the statement is correct in this case). If f has (at least) a root α_1, then according to the Factor Theorem (see Lemma 22) we have

$$f(x)=(x-\alpha_1)\cdot f_2(x)$$

for some polynomial $f_2(x)$. Now it may happen that $f_2(x)$ has no roots (and then $f(x)$ has exactly one root, so the statement is correct because $1\leq n$). But if $f_2(x)$ does have

a root α_2 (possibly equal to α_1), then

$$f_2(x) = (x - \alpha_2) \cdot f_3(x),$$

whence

$$f(x) = (x - \alpha_1)(x - \alpha_2) \cdot f_3(x).$$

Continuing in this way, sooner or later we arrive at

$$f(x) = (x - \alpha_1) \cdots (x - \alpha_m) \cdot f_{m+1}(x),$$

where $f_{m+1}(x)$ has no roots in F (possibly because it is a constant). In fact, this must occur for some $m \leq \deg(f) = n$, because taking degrees in the above equality we find $\deg(f) = m + \deg(f_{m+1}) \geq m$.

Now let β be a root of $f(x)$. To complete the proof it will be enough to show that β equals one among $\alpha_1, \ldots, \alpha_m$, of which there are at most m distinct ones (possibly fewer!), and consequently at most n as we have just shown $m \leq n$. In fact, if $f(\beta) = 0$ then

$$0 = f(\beta) = (\beta - \alpha_1) \cdots (\beta - \alpha_m) \cdot f_{m+1}(\beta).$$

We have $f_{m+1}(\beta) \neq 0$ because $f_{m+1}(x)$ has no roots in F, and so at least one other factor in the above product must vanish, say $\beta - \alpha_j$, and hence $\beta = \alpha_j$, as desired. □

In particular, any nonzero polynomial has finitely many roots. This behaviour of polynomial functions is different from other familiar functions, for example the function $f(x) = \sin(x)$ has infinitely many real zeroes, namely,

$$\sin(x) = 0 \qquad \Leftrightarrow \qquad x = 2\pi k \quad \text{for} \quad k \in \mathbb{Z}.$$

Corollary 29. *A polynomial $f(x)$ of degree n is uniquely determined by the values it takes on $n + 1$ distinct elements of F.*

Proof. We are assuming that for $n + 1$ distinct numbers $b_1, \ldots, b_{n+1}$ we know the values

$$f(b_1) = c_1, \quad f(b_2) = c_2, \quad \ldots \quad f(b_{n+1}) = c_{n+1}.$$

Suppose $g(x)$ is any polynomial of degree n which satisfies

$$g(b_1) = c_1, \quad g(b_2) = c_2, \quad \ldots \quad g(b_{n+1}) = c_{n+1}.$$

Then the difference $h(x) = f(x) - g(x)$ is either zero or a nonzero polynomial of degree *at most* n, and it satisfies

$$h(b_1) = 0, \quad h(b_2) = 0, \quad \ldots \quad h(b_{n+1}) = 0.$$

Hence $h(x)$ has at least $n + 1$ roots, while a nonzero polynomial of degree at most n has at most n roots. Consequently, $h(x)$ can only be the zero polynomial, and hence $g(x) = f(x)$. □

An important consequence of the above corollary is that two different polynomials $f(x), g(x) \in \mathbb{R}[x]$ give rise to different functions $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$. Consequently, the apparently simpler alternate definition of a polynomial as a function $f\colon \mathbb{R} \to \mathbb{R}$ of a particular shape (that is, which can be written as $f(x) = a_n x^n + \cdots + a_1 x + a_0$) is actually equivalent to the one we are using (so we can identify polynomials with the functions they gives rise to), but only *if we work over an infinite field,* such as $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$.

Example. Consider the polynomials $f(x) = \bar{1}x = x$ and $g(x) = \bar{1}x^2 = x^2$, where the coefficients belong to the field of two elements $\mathbb{F}_2$. According to our definition of polynomials they are different polynomials (because their coefficients are different; and actually even their degrees) but give rise to the same function $\mathbb{F}_2 \to \mathbb{F}_2$, because

$$f(\bar{0}) = \bar{0} = g(\bar{0}), \qquad \text{and} \qquad f(\bar{1}) = \bar{1} = g(\bar{1}).$$

19. Polynomial interpolation

Given real numbers $\alpha_1, \alpha_2, \beta_1, \beta_2$, with $\alpha_1 \neq \alpha_2$, there is a unique polynomial $f(x)$ of degree at most one (hence a linear polynomial or a constant polynomial), such that

$$f(\alpha_1) = \beta_1, \quad \text{and} \quad f(\alpha_2) = \beta_2.$$

In fact, because $f(x) = ax + b$ for some $a, b \in \mathbb{R}$, the required conditions amount to

$$\begin{cases} a \cdot \alpha_1 + b = \beta_1 \\ a \cdot \alpha_2 + b = \beta_2 \end{cases}$$

Solving this system we would find

$$a = \frac{\beta_2 - \beta_1}{\alpha_2 - \alpha_1}, \quad \text{and then} \quad b = \beta_1 - \frac{\beta_2 - \beta_1}{\alpha_2 - \alpha_1},$$

and hence there is a unique solution, which can be written as

$$f(x) = \frac{\beta_2 - \beta_1}{\alpha_2 - \alpha_1} \cdot (x - \alpha_1) + \beta_1.$$

The following result generalises this fact to an arbitrary number of values. We state it for complex numbers rather than real numbers, but we could as well take for $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ elements of any field (the same field for all of them).

Theorem 30 (Interpolation theorem). *Given n distinct complex numbers $\alpha_1, \ldots, \alpha_n$, and n arbitrary (not necessarily distinct) complex numbers $\beta_1, \ldots, \beta_n$, there is a unique polynomial $f(x)$ of degree less than n such that*

$$f(\alpha_1) = \beta_1, \quad \ldots, \quad f(\alpha_n) = \beta_n.$$

(Optional) Proof. We have already proved the uniqueness of $f(x)$ in Corollary 29. There are several ways to prove the existence of $f(x)$, both direct and indirect (and hence not explicit). We present an explicit proof, based on *Lagrange interpolation.*

The polynomial

$$(x-\alpha_2)(x-\alpha_3)\cdots(x-\alpha_n) = \frac{(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n)}{x-\alpha_1}$$

has $\alpha_2, \alpha_3, \ldots, \alpha_n$ as roots, and so it takes the value 0 on $\alpha_2, \alpha_3, \ldots, \alpha_n$. On α_1 it takes the value $(\alpha_1-\alpha_2)(\alpha_1-\alpha_3)\cdots(\alpha_1-\alpha_n)$. Consequently, the polynomial

$$p_1(x) = \frac{(x-\alpha_2)(x-\alpha_3)\cdots(x-\alpha_n)}{(\alpha_1-\alpha_2)(\alpha_1-\alpha_3)\cdots(\alpha_1-\alpha_n)}$$

satisfies

$$p_1(\alpha_1) = 1, \quad p_1(\alpha_2) = 0, \quad \ldots, \quad p_1(\alpha_n) = 0.$$

Similarly, we can construct a polynomial

$$p_1(x) = \frac{(x-\alpha_1) \quad (x-\alpha_3)(x-\alpha_4)\cdots(x-\alpha_n)}{(\alpha_2-\alpha_1) \quad (\alpha_2-\alpha_3)(\alpha_2-\alpha_4)\cdots(\alpha_2-\alpha_n)}$$

(where $x-\alpha_2$ is omitted from the numerator, and $\alpha_2-\alpha_2$ is correspondingly omitted from the denominator), which takes the value 1 on α_2 and vanishes (that means, it takes the value zero) on $\alpha_1, \alpha_3, \alpha_4, \ldots, \alpha_n$. After finding $p_3(x), \ldots, p_n(x)$ with similar properties, we see that the polynomial

$$f(x) = \beta_1 \cdot p_1(x) + \cdots + \beta_1 \cdot p_1(x)$$

satisfies

$$f(\alpha_1) = \beta_1, \quad \ldots, \quad f(\alpha_n) = \beta_n,$$

as required, and has degree at most $n-1$ because each of $p_1(x), \ldots, p_n(x)$ has degree exactly $n-1$. □

Each polynomial $p_k(x)$ in the above proof is obtained as

$$p_k(x) = g_k(x)/g_k(\alpha_k), \quad \text{where} \quad g_k(x) = \prod_{j\neq k}(x-\alpha_j).$$

Note that it may be quicker to compute $(x-\alpha_1)\cdots(x-\alpha_n)$ first, a polynomial of degree n, and then obtain each $g_k(x)$ by dividing that by $x-\alpha_k$, which can be efficiently done using Ruffini's rule. [8]

[8] A rough way to see why is noting that computing $(x-\alpha_1)\cdots(x-\alpha_n)$ requires $n-1$ polynomial multiplications, and on the way to the conclusion we would have already computed one of the $g_k(x)$, say $g_n(x) = (x-\alpha_1)\cdots(x-\alpha_{n-1})$. Then it remains to obtain the remaining $g_k(x)$ by applying Ruffini's rule $n-1$ times. This total of $2n-2$ multiplications/divisions should be contrasted with the $n(n-1)$ multiplications required to compute each $g_k(x)$ separately. (More precise computational estimates would take into account that polynomial multiplications may take different times depending on the size of the factors, but then we would still observe a similar difference in speed of the two methods.)

EXAMPLE. Find the unique polynomial $f(x)$ of degree at most three, such that

$$f(-2) = -5, \qquad f(-1) = 3, \qquad f(1) = 1, \qquad f(2) = 3.$$

The polynomial

$$p_1(x) = \frac{(x+1)(x-1)(x-2)}{(-2+1)(-2-1)(-2-2)} = -\frac{1}{12}(x^3 - 2x^2 - x + 2)$$

has -1, 1 and 2 as roots and satisfies $p_1(-2) = 1$. The polynomial

$$p_2(x) = \frac{(x+2)(x-1)(x-2)}{(-1+2)(-1-1)(-1-2)} = \frac{1}{6}(x^3 - x^2 - 4x + 4)$$

has -2, 1 and 2 as roots and satisfies $p_2(-1) = 1$. The polynomial

$$p_3(x) = \frac{(x+2)(x+1)(x-2)}{(1+2)(1+1)(1-2)} = -\frac{1}{6}(x^3 + x^2 - 4x - 4)$$

has -2, -1 and 2 as roots and satisfies $p_3(1) = 1$. The polynomial

$$p_4(x) = \frac{(x+2)(x+1)(x-1)}{(2+2)(2+1)(2-1)} = \frac{1}{12}(x^3 + 2x^2 - x - 2)$$

has -2, -1 and 1 as roots and satisfies $p_4(2) = 1$.

Note that because of the particular symmetry of our problem, $p_3(x)$ and $p_4(x)$ could have also been obtained as $p_3(x) = p_2(-x)$ and $p_4(x) = p_1(-x)$. Of course it will not be so in general.

We conclude that

$$\begin{aligned} f(x) &= -5 \cdot p_1(x) + 3 \cdot p_2(x) + p_3(x) + 3 \cdot p_4(x) \\ &= \frac{5}{12}x^3 - \frac{5}{6}x^2 - \frac{5}{12}x + \frac{5}{6} \\ &\quad + \frac{1}{2}x^3 - \frac{1}{2}x^2 - 2x + 2 \\ &\quad - \frac{1}{6}x^3 - \frac{1}{6}x^2 + \frac{2}{3}x + \frac{2}{3} \\ &\quad + \frac{1}{4}x^3 + \frac{1}{2}x^2 - \frac{1}{4}x - \frac{1}{2} \\ &= x^3 - x^2 - 2x + 3. \end{aligned}$$

EXAMPLE. Find the unique polynomial $g(x)$ of degree at most three, such that

$$g(-2) = 1, \qquad g(-1) = -1, \qquad g(1) = -1, \qquad g(2) = 1.$$

We can reuse the calculations of the previous example, and find

$$f(x) = p_1(x) - p_2(x) - p_3(x) + p_4(x) = \frac{2}{3}x^2 - \frac{5}{3}.$$

Hence this time the required polynomial has actually degree two. According to the interpolation theorem, Theorem 30, it is the unique polynomial of degree *less than four,* which is the same as *at most three,* which satisfies the given conditions. Of course, had

we known that it has degree at most two, it would have been uniquely determined by any three of those four conditions, again according to the interpolation theorem.

20. Irreducibility and roots for quadratic and cubic polynomials

Consider a polynomial $f(x)$, of positive degree, with coefficients in a field F. Recall that, according to the Factor Theorem an element α of F is a root of $f(x)$ (that is, $f(\alpha) = 0$) exactly when $x - \alpha$ is a factor of $f(x)$ (that is, it divides $f(x)$). Hence each time we find a root α of $f(x)$ we have achieved a partial factorisation of $f(x)$ as $f(x) = (x - \alpha)\, g(x)$, for some polynomial $g(x)$ (again with coefficients in F).

In particular, if a polynomial $f(x)$ of degree larger than one has a root in F, then it is reducible in $F[x]$. For polynomials of degree two or three this implication can be inverted.

PROPOSITION 31. *A quadratic or cubic polynomial (that is, of degree two or three) over a field F is irreducible over F exactly when it does not have any root in F.*

PROOF. The statement is equivalent to the following: a quadratic or cubic polynomial over a field F is reducible over F exactly when it has some root in F (meaning *at least one root in F*). Now we prove this statement.

If our polynomial, say $f(x)$, has a root α in F, then according to the Factor Theorem we have $f(x) = (x-\alpha)\, g(x)$ for some polynomial $g(x) \in F[x]$, and hence $f(x)$ is reducible over F.

Conversely, if $f(x)$ is reducible over F, then $f(x) = g(x)\, h(x)$ for some polynomials of positive degree $g(x), h(x) \in F[x]$. Because $\deg\big(f(x)\big) = \deg\big(g(x)\big) + \deg\big(h(x)\big)$, and $\deg\big(f(x)\big)$ equals two or three, then at least one of the factors, say $g(x)$, must have degree one, and hence be of the form $g(x) = ax + b$, with $a, b \in F$ and $a \neq 0$. Then $-b/a$ is a root of $g(x)$, and hence of $f(x)$. In conclusion, $f(g)$ has at least one root in F. □

Of course a polynomial of degree one, hence of the form $ax + b$ with $a, b \in F$ and $a \neq 0$, is irreducible but has always a root in F, namely $-b/a$. This criterion for being irreducible (or reducible) does not work for polynomials of degree four or higher. In fact, it is possible that a polynomial (of degree at least four) has a proper factorisation over F even if it does not have any root in F, as the following examples show.

EXAMPLE. The polynomial $x^4 + 5x^2 + 4$ factorises over $\mathbb{R}$ as $(x^2 + 1)(x^2 + 4)$, but it has no real roots. In fact, its complex roots are $\pm i$ and $\pm 2i$, but none of them is real. Hence $x^4 + 5x^2 + 4$ is reducible over $\mathbb{R}$ (and, in fact, it is the product of the two irreducible polynomials $x^2 + 1$ and $x^2 + 4$), despite having no roots in $\mathbb{R}$.

EXAMPLE. The polynomial $x^4 + 1$ has no roots in $\mathbb{R}$, but is not irreducible in $\mathbb{R}[x]$. More generally, we have

$$\begin{aligned} x^4 + a^4 &= (x^4 + 2a^2x^2 + a^4) - 2a^2x^2 \\ &= (x^2 + a^2)^2 - (\sqrt{2}ax)^2 \\ &= [(x^2 + a^2) - \sqrt{2}ax][(x^2 + a^2) + \sqrt{2}ax] \\ &= (x^2 - \sqrt{2}ax + a^2)(x^2 + \sqrt{2}ax + a^2). \end{aligned}$$

Hence if $a \neq 0$ is a real number, then $x^4 + a^4$ has no real roots, but is reducible in $\mathbb{R}[x]$, and the factorisation given here is its complete factorisation in $\mathbb{R}[x]$. For example, $x^4 + 4 = (x^2 - 2x + 2)(x^2 + 2x + 2)$, a factorisation which is even in $\mathbb{Q}[x]$.

21. The Fundamental Theorem of Algebra

We know how to solve equations of degree 1 and 2. In case of quadratic equations the formula involves taking a square root (of the discriminant). More complicated formulas exist for finding the roots of polynomials of degree three and four (which means solving the corresponding equations). Those formulas (due to Del Ferro/Tartaglia/Cardano for cubics, and Ferrari/Cardano for quartics, all in 16th century) involve the usual algebraic operations together with taking square and cube roots.

However, the Norwegian mathematician Abel proved in 1824 (building on previous partial work of Ruffini) that there is no analogous formula expressing the roots of a polynomial of degree five or higher in the general case (the Abel-Ruffini Theorem). This essentially means that there are polynomials $f(x)$ of degree five (for example $x^5 - x - 1$) whose roots cannot be described by taking the coefficients of $f(x)$ and manipulating them by the usual algebraic operations together with the operations of taking nth roots (forming radicals), in the way we do for quadratic polynomials (and can be done for cubic and quartic polynomials).

Despite the impossibility of a general formula for degree larger than four, the fundamental Theorem of Algebra (first proof by Argand in 1806, more proofs by Gauss soon later) asserts that at least one complex root exixts for any non-constant polynomial.

THEOREM 32 (Fundamental Theorem of Algebra). *Every polynomial in $\mathbb{C}[x]$ of positive degree has at least one root in $\mathbb{C}$.*

COROLLARY 33. *The irreducible polynomials in $\mathbb{C}[x]$ are those of degree one.*

Many proofs are known but none is really easy, so we will not prove the theorem.

PROOF. The polynomials of degree one are always irreducible, so we only need to prove the converse: if $f(x)$ is irreducible in $\mathbb{C}[x]$, then $f(x)$ has degree one.

Because of the Fundamental Theorem of Algebra, $f(x)$ has at least one root α. By the Factor Theorem we have $f(x) = (x - \alpha)\, g(x)$, for some $g(x) \in \mathbb{C}[x]$. This cannot be a proper factorisation of $f(x)$, because $f(x)$ is irreducible, and so $g(x)$ must be a constant. Consequently, $f(x)$ equals $x - \alpha$ times a nonzero constant, and so $f(x)$ has degree one. □

COROLLARY 34. *Every polynomial of positive degree in $\mathbb{C}[x]$ is a product of polynomials of degree one (also called linear polynomials).*

PROOF. We know that if F is any field then every polynomial of positive degree in $F[x]$ is a product of irreducible polynomials. But when $F = \mathbb{C}$ all irreducible polynomials have degree one. □

EXAMPLE. One can show that the polynomial $x^5 - x - 1$ is irreducible in $\mathbb{Q}[x]$. There exists no formula for the roots (using only algebraic operations and radicals), but one can find numerically that one root is approximately 1.167 (and more precisely 1.167303978...). This is a real root, but of course, in particular, it is a complex root. According to the Factor Theorem, $x - 1.167$ (approximately) divides $x^5 - x - 1$, and Ruffini's rule gives us [9]

	1	0	0	0	−1	−1
1.167		1.167	1.362	1.590	1.856	−1
	1	1.167	1.362	1.590	0.856	0

and so we find

$$x^5 - x - 1 \approx (x - 1.167)(x^4 + 1.167x^3 + 1.362x^2 + 1.590x + 0.856).$$

In turn, the factor of degree 4 has at least one complex root, and continuing in this way we eventually find the complete complex factorisation of $x^5 - x - 1$,

$$\approx (x-1.167)(x-0.181+1.083\,i)(x-0.181-1.083\,i)(x+0.764+0.352\,i)(x+0.764-0.352\,i).$$

We see that the non-real roots come in conjugate pairs (see next section), and if we multiply together the corresponding factors we find the complete factorisation of x^5-x-1 in $\mathbb{R}[x]$, which is

$$x^5 - x - 1 \approx (x - 1.167)(x^2 - 0.362x + 1.207)(x^2 + 1.529x + 0.709).$$

We will see the theory of the factorisation over $\mathbb{R}$ in the next section.

[9] Here we have written only three decimal digits after the point of each number, but we have done the calculations with higher precision. In reality we will never find remainder exactly zero with a calculator, as we are using an approximation of the true root. For example, $1.167^5 - 1.167 - 1 \approx -0.0025$.

22. Roots and factorisations of a polynomial with real coefficients

Recall that any complex number α can be uniquely written in the form $\alpha = s + it$, where s and t are real numbers. For now we may actually take this as a definition of complex numbers, and define addition and multiplication by treating them like ordinary "expressions" except that every time we encounter i^2 we may replace it with -1 (so i is not a letter like any other but satisfies the "simplification rule" $i^2 = -1$). Hence complex numbers can be added, subtracted, and multiplied as follows:

$$(s + it) \pm (u + iv) = (s \pm u) + i(t \pm v),$$
$$(s + it)(u + iv) = su + i(sv + tu) + i^2tv = (su - tv) + i(sv + tu).$$

To perform division it is useful to introduce the *conjugate* of a complex number $\alpha = s + it$, which is $\bar{\alpha} = s - it$. Because $\alpha\bar{\alpha} = (s + it)(s - it) = s^2 + t^2 = |\alpha|^2$ (where the *modulus* $|\alpha|$ is the nonnegative real number given by $|s + it| = \sqrt{s^2 + t^2}$), the reciprocal of α can be computed as $\dfrac{1}{\alpha} = \dfrac{\bar{\alpha}}{|\alpha|^2}$, and general division of complex numbers then easily follows.

Now note that complex conjugation has the following properties, which hold for any $\alpha, \beta \in \mathbb{C}$:

$$\overline{\alpha + \beta} = \overline{\alpha} + \overline{\beta}, \qquad \overline{\alpha\beta} = \overline{\alpha}\,\overline{\beta}.$$

These properties essentially say that conjugation, in the sense of the map $\mathbb{C} \to \mathbb{C}$ taking $\alpha \mapsto \overline{\alpha}$, is an *automorphism* of $\mathbb{C}$. They can be verified by writing $\alpha = s+it$ and $\beta = u+iv$ and checking

$$\overline{(s + it) + (u + iv)} = \overline{(s + u) + i(t + v)} = (s + u) - i(t + v) = (s - it) + (u - iv),$$
$$\overline{(s + it)(u + iv)} = \overline{(su - tv) + i(sv + tu)} = (su - tv) - i(sv + tu) = (s - it)(u - iv).$$

Note also that a complex number α is actually real precisely when $\overline{\alpha} = \alpha$. Other properties follow, such as $\overline{\alpha - \beta} = \overline{\alpha + (-1)\beta} = \overline{\alpha} + \overline{(-1)\beta} = \overline{\alpha} + (-1)\overline{\beta} = \overline{\alpha} - \overline{\beta}$ for subtraction, and a similar one for division, $\overline{\alpha/\beta} = \overline{\alpha}/\overline{\beta}$. Also, $\overline{\alpha^2} = \overline{\alpha}^{\,2}$, and more generally $\overline{\alpha^n} = \overline{\alpha}^{\,n}$.

As an application of these basic properties of conjugation, we now show that if a complex number is a root of a polynomial with real coefficients, then its conjugate is also a root.

LEMMA 35. *If a complex number $\alpha = s + it$ is a root of a polynomial $f(x) \in \mathbb{R}[x]$, then its conjugate $\overline{\alpha} = s - it$ is a root as well.*

PROOF. Write the polynomial as $f(x) = a_n x^n + \cdots + a_2 x^2 + a_1 x + a_0$, hence $a_j \in \mathbb{R}$. Then if α is any complex number, not necessarily a root of $f(x)$, we have

$$\begin{aligned}
f(\overline{\alpha}) &= a_n\,\overline{\alpha}^{\,n} + \cdots + a_2\,\overline{\alpha}^{\,2} + a_1\,\overline{\alpha} + a_0 \\
&= a_n\,\overline{\alpha^n} + \cdots + a_2\,\overline{\alpha^2} + a_1\,\overline{\alpha} + a_0 \qquad \text{(because } \overline{\alpha^n} = \overline{\alpha}^{\,n}\text{)} \\
&= \overline{a_n\,\alpha^n} + \cdots + \overline{a_2\,\alpha^2} + \overline{a_1\,\alpha} + \overline{a_0} \qquad \text{(because } \overline{\alpha\beta} = \overline{\alpha}\,\overline{\beta} \text{ and } \overline{a_j} = a_j\text{)} \\
&= \overline{a_n\,\alpha^n + \cdots + a_2\,\alpha^2 + a_1\,\alpha + a_0} \qquad \text{(because } \overline{\alpha+\beta} = \overline{\alpha} + \overline{\beta}\text{)} \\
&= \overline{f(\alpha)}.
\end{aligned}$$

In particular, if α is a root of $f(x)$, which means $f(\alpha) = 0$, then $f(\overline{\alpha}) = \overline{f(\alpha)} = 0$, and so $\overline{\alpha}$ is also a root. □

Note that if α is any complex number, then the polynomial

$$(x - \alpha)(x - \overline{\alpha}) = x^2 - (\alpha + \overline{\alpha})x + \alpha\overline{\alpha}$$

has real coefficients. In fact, if $\alpha = r + it$, then $\alpha + \overline{\alpha} = (r + it) + (r - it) = 2r$, and $|\alpha|^2 = \alpha\overline{\alpha} = (r + it)(r - it) = r^2 + t^2$. That quadratic polynomial has discriminant

$$(\alpha + \overline{\alpha})^2 - 4\alpha\overline{\alpha} = (\alpha - \overline{\alpha})^2 = (2it)^2 = -4t^2,$$

which is zero if α is real (and it must be so because α is a double root in that case), but is negative otherwise (and it must be so because the polynomial has no real roots in that case).

THEOREM 36. *The irreducible polynomials in $\mathbb{R}[x]$ are precisely the polynomials of degree 1, and the quadratic polynomials $ax^2 + bx + c$ with $b^2 - 4ac < 0$.*

PROOF. A polynomial of degree 1 is always irreducible. We have just seen that a polynomial $ax^2 + bx + c$ with $b^2 - 4ac < 0$ has no real roots. We know that for polynomials of degree two or three this implies that the polynomial is irreducible.

Now we prove the converse. Let $f(x)$ be any irreducible polynomial in $\mathbb{R}[x]$. By definition, it must have positive degree, and so by the Fundamental Theorem of Algebra it has at least one complex root α. By the Factor Theorem (in $\mathbb{C}[x]$) we have $f(x) = (x - \alpha)\,g(x)$, for some $g(x) \in \mathbb{C}[x]$.

If α is real, then $g(x) \in \mathbb{R}[x]$. Because we are assuming $f(x)$ irreducible, it follows that $g(x)$ is constant, and so $f(x)$ has degree 1.

If α is not real, then we have seen that $\overline{\alpha}$ is also a root of $f(x) = (x - \alpha)\,g(x)$, and being different from α it must be a root of $g(x)$, so $g(\overline{\alpha}) = 0$. Hence $x - \overline{\alpha}$ must divide $g(x)$, and so

$$f(x) = (x - \alpha)(x - \overline{\alpha}) \cdot h(x) = [x^2 - (\alpha + \overline{\alpha})x + \alpha\overline{\alpha}] \cdot h(x)$$

for some $h(x) \in \mathbb{C}[x]$. But we have seen earlier that the quadratic factor $x^2-(\alpha+\overline{\alpha})x+\alpha\overline{\alpha}$ has real coefficients, and so $h(x)$ has real coefficients as well. Because we are assuming $f(x)$ irreducible, it follows that $g(x)$ is constant, and so $f(x)$ has degree 2. Also, it has negative discriminant as claimed in the theorem because its roots α and $\overline{\alpha}$ are not real. □

COROLLARY 37. *Every polynomial of positive degree in $\mathbb{R}[x]$ is a product of polynomials of degree one and of quadratic polynomials with negative discriminant.*

Consequently, a polynomial of odd degree in $\mathbb{R}[x]$ has always at least one real root.

EXAMPLE. Consider the polynomial $f(x) = 4x^4+20x^3+30x^2-40x+26$, and suppose that we have somehow found out that $-3+2i$ is a root, meaning that $f(-3+2i) = 0$. Our task is to find the remaining complex roots, and obtain a complete factorisation of $f(x)$ in $\mathbb{C}[x]$.

According to the Factor Theorem, $x+3-2i$ is a factor of $f(x)$. Hence we divide $f(x)$ by $x+3-2i$ using Ruffini's rule:

	4	20	30	-40	26
$-3+2i$		$-12+8i$	$-40-8i$	$46+4i$	-26
	4	$8+8i$	$-10-8i$	$6+4i$	0

This confirms that $f(-3+2i)$ is actually a root of $f(x)$ as claimed. It also tells us that

$$f(x) = (x+3-2i)\cdot[4x^3+(8+8i)x^2+(-10-8i)x+(6+4i)].$$

Because $-3+2i$ is a root of $f(x)$ we know that its conjugate $-3-2i$ is a root as well, and because that is not a root of its factor $x+3-2i$ it must be a root of the other factor, a cubic polynomial. Dividing that by $x+3+2i$ using Ruffini's rule, we find

	4	$8+8i$	$-10-8i$	$6+4i$
$-3-2i$		$-12-8i$	$12+8i$	$-6-4i$
	4	-4	2	0

Hence $f(x) = (x+3-2i)(x+3+2i)(4x^2-4x+2)$. Finally, the roots of the quadratic factor can easily be found to be $(1 \pm i)/2$, by means of the usual formula, and so the complete factorisation of $f(x)$ in $\mathbb{C}[x]$ is

$$f(x) = (x+3-2i)(x+3+2i)(2x-1-i)(2x-1+i).$$

The complete factorisation of $f(x)$ in $\mathbb{R}[x]$ can be found by multiplying together the pairs of linear factors corresponding to conjugate roots: $f(x) = (x^2+6x+13)(4x^2-4x+2)$.

Lecture notes of Algebra. Week 6

23. Rational roots of a polynomial with integer coefficients

There is a test which allows, with a finite amount of calculations, to find *all* rational roots of a polynomials with rational coefficients, or to conclude that none exists if none is found. It is a rather specialised test, but it is sometimes useful, and its proof is a good illustration of the use of Arithmetical Lemma B on divisibility.

Note that Arithmetical Lemma B can be applied repeatedly to a product of more than two factors, and hence: if an integer divides a product of several integers, and is coprime (separately) with each of the factors except one, then it divides that factor. For example, if a divides a product bcd, and $(a, b) = 1$ and $(a, c) = 1$, then a must divide d. In fact, because a divides $b(cd)$, and $(a, b) = 1$ we have that a divides cd, and then because $(a, c) = 1$ we have that a divides d.

Before we apply the test we we may reduce to the case of a polynomial with integer coefficients by multiplying our polynomial by a suitable integer (say the least common multiple of all denominators occurring in the coefficients). Then we may divide by any common factor of the coefficients; strictly speaking, this is not required for the validity of the following test, but it may avoid us lots of superfluous calculations.

THEOREM 38 (The Rational Root Test). *Consider a polynomial* $f(x) = a_n x^n + \cdots + a_1 x + a_0$ *with integer coefficients and* $a_n a_0 \neq 0$. *If* r/s *is a rational root of* $f(x)$, *written as a fraction of integers in lowest terms (that is, with* $(r, s) = 1$*), then* r *divides the constant term* a_0, *and* s *divides the leading coefficient* a_n.

PROOF. After expanding $s^n \cdot f(r/s) = 0$ we find

$$a_n r^n + a_{n-1} r^{n-1} s + a_{n-2} r^{n-2} s^2 + \cdots + a_2 r^2 s^{n-2} + a_1 r s^{n-1} + a_0 s^n = 0.$$

Because r divides all terms preceding the last one, it must divide the last term as well, that is, $r \mid a_0 s^n$. But because $(r, s) = 1$, Arithmetical Lemma B implies that r divides a_0.

In a similar way, because s divides all terms following the first one, it must divide the first term $a_n r^n$ as well. Because $(r, s) = 1$ it follows that s divides a_n. □

EXAMPLE. Find all the rational roots of the polynomial $f(x) = 2x^3 + 15x^2 + 27x + 10$, and then factorise it over $\mathbb{Q}$. According to the test, if $r/s \in \mathbb{Q}$ is a root of $f(x)$, with $\gcd(r, s) = 1$, then r divides 10 and s divides 2, hence $r \in \{\pm 1, \pm 2, \pm 5, \pm 10\}$ and $s \in \{\pm 1, \pm 2\}$. Consequently, the possibilities for r/s are

$$\pm 1,\ \pm 2,\ \pm 5,\ \pm 10,\ \pm\frac{1}{2},\ \pm\frac{5}{2}.$$

However, because the coefficients of the polynomial are all positive, no positive real number can be a root, and so we only have to test the negative ones. Going through the list from left to right, we find $f(-2) = f(-5) = f(-1/2) = 0$, at which point we can stop because $f(x)$ cannot have more than three roots, and we conclude that

$$f(x) = 2(x+2)(x+5)\big(x+\tfrac{1}{2}\big) = (x+2)(x+5)(2x+1),$$

which is the desired complete factorisation of $f(x)$ over $\mathbb{Q}$. The last expression is actually a complete factorisation over $\mathbb{Z}$. (It is a general fact, known as *Gauss' lemma,* that any factorisation over $\mathbb{Q}$ of a polynomial with integer coefficients leads to a corresponding factorisation over $\mathbb{Z}$ by suitably rearranging some scalar factors.) Alternatively, once we have found the first root -2 we may divide $f(x)$ by $x+2$, and then proceed to factorise the resulting quadratic polynomial.

We can use the Rational Root Test to prove that certain radicals represent irrational numbers, as follows.

EXAMPLE. We prove that $\sqrt{3}$ is irrational. We start with noting that $\sqrt{3}$ is a root of the polynomial $x^2 - 3$. By the Rational Root Test, if r/s is a rational root of $x^2 - 3$, with $r, s \in \mathbb{Z}$ and $(r, s) = 1$, then $r \mid 3$ and $s \mid 1$, and so $r/s \in \{\pm 1, \pm 3\}$. None of those numbers is a root, hence $x^2 - 3$ has no rational root, and so $\sqrt{3}$ is irrational.

EXAMPLE. We prove that $\sqrt[3]{25/3}$ is irrational. Here $\sqrt[3]{25/3}$ is a root of the polynomial $3x^3 - 25$. By the Rational Root Test, if r/s is a rational root of $3x^3 - 25$, with $r, s \in \mathbb{Z}$ and $(r, s) = 1$, then r divides 25 and s divides 3, and so $r/s \in \{\pm 1, \pm 5, \pm 25, \pm 1/3, \pm 5/3, \pm 25/3\}$.

To conclude that $\sqrt[3]{25/3}$ cannot be rational it is enough to check that none of those 12 rational numbers is a root of the polynomial (and so $\sqrt[3]{25/3}$ cannot be equal to any of those rational numbers). However, it may not be necessary to check them all. For example, no negative real number can possibly be a root of $3x^3 - 25$, and so we only need to check the positive ones. We can do even better if we are able to locate the real roots of the polynomial more precisely. For example, noting that $3 \cdot 2^3 - 25 = -1 < 0$, and $3 \cdot 3^3 - 25 = 56 > 0$, and that the function $x \mapsto x^3$ is increasing, any real root of $3x^3 - 25$ must be larger than 2 and less than 3. However, none of the candidates which we have found for rational roots is between 2 and 3, and so we conclude that $3x^2 - 25$ has no rational root, and hence that $\sqrt[3]{25/3}$ is irrational.

24. Some special polynomials: biquadratic polynomials

A *biquadratic polynomial* is a polynomial of degree four where the terms of odd degree are missing, and so has the form $ax^4 + bx^2 + c$, with $a \neq 0$. Because it can be viewed as $a(x^2)^2 + bx^2 + c$, the standard way of finding its roots is setting $y = x^2$, and then solving $ay^2 + by + c = 0$ (by completing the square, or by the explicit formula). If β_1, β_2 are the

roots of this quadratic equation in y, then the roots of the biquadratic polynomial are the solutions of either $x^2 = \beta_1$ or $x^2 = \beta_2$, and so they are the square roots of β_1 and the square roots of β_2.

EXAMPLE. To find the roots of the biquadratic polynomial $2x^4 + x^2 - 6$ we set $x^2 = y$ and then calculate the roots of the resulting quadratic polynomial $2y^2 + y - 6$, finding $y = 3/2$ or $y = -2$. In terms of x this means $x^2 = 3/2$ or $x^2 = -2$, which leads to $x = \pm\sqrt{3/2} = \pm\sqrt{6}/2$ or $x = \pm i\sqrt{2}$. Hence the full factorisation of the polynomial over $\mathbb{C}$ is

$$\begin{aligned} 2x^4 + x^2 - 6 &= 2(x - \sqrt{6}/2)(x + \sqrt{6}/2)(x - i\sqrt{2})(x + i\sqrt{2}) \\ &= (\sqrt{2}\,x - \sqrt{3})(\sqrt{2}\,x + \sqrt{3})(x - i\sqrt{2})(x + i\sqrt{2}), \end{aligned}$$

whichever form we prefer (as the latter has no denominators, but more radicals). Its complete factorisation over $\mathbb{R}$ is

$$2x^4 + x^2 - 6 = 2(x - \sqrt{6}/2)(x + \sqrt{6}/2)(x^2 + 2),$$

and $x^2 + 2$ is irreducible over $\mathbb{R}$ because it has degree two and has no real roots. Finally, its complete factorisation over $\mathbb{Q}$ is

$$2x^4 + x^2 - 6 = (2x^2 - 3)(x^2 + 2),$$

where again the two quadratic factors are irreducible over $\mathbb{Q}$ because they have no rational roots.

25. (Optional) Double radicals

A *double radical* is an expression of the form $\sqrt{a \pm \sqrt{b}}$. (This is a special case of a *nested radical*, see `https://en.wikipedia.org/wiki/Nested_radical`.) Such an expression may occur, for example, when solving biquadratic equations and quartic self-reciprocal equations, or already when solving quadratic equations if the coefficients involve radicals. A double radical can sometimes be expressed as a sum of difference of *simple radicals*, using the identity

(Double Radical Identity)
$$\sqrt{a \pm \sqrt{b}} = \sqrt{\frac{a + \sqrt{a^2 - b}}{2}} \pm \sqrt{\frac{a - \sqrt{a^2 - b}}{2}}.$$

Of course the right-hand side is the sum or difference of two simple radicals only when $a^2 - b$ is an exact square, otherwise the identity expresses a double radical as a more complicated expression, a sum of two double radicals.

EXAMPLE. We have

$$\sqrt{5 \pm 2\sqrt{6}} = \sqrt{5 \pm \sqrt{24}} = \sqrt{\frac{5 + \sqrt{5^2 - 24}}{2}} \pm \sqrt{\frac{5 - \sqrt{5^2 - 24}}{2}} = \sqrt{3} \pm \sqrt{2}.$$

In fact, squaring $\sqrt{3} \pm \sqrt{2}$ we get

$$(\sqrt{3} \pm \sqrt{2})^2 = \sqrt{3}^2 \pm 2\sqrt{3}\sqrt{2} + \sqrt{2}^2 = 3 \pm 2\sqrt{6} + 2 = 5 \pm 2\sqrt{6}.$$

EXAMPLE. We have

$$\sqrt{6 \pm 2\sqrt{3}} = \sqrt{6 \pm \sqrt{12}} = \sqrt{\frac{6 + \sqrt{6^2 - 12}}{2}} \pm \sqrt{\frac{5 - \sqrt{6^2 - 12}}{2}} = \sqrt{3 + \sqrt{6}} \pm \sqrt{3 - \sqrt{6}}.$$

This is correct but not very useful.

For the Double Radical Identity to really make sense we should put proper limitations on the values allowed for a and b. Convenient limitations for the present exposition are that a, b are real numbers with $a \geq 0$ and $0 \leq b \leq a^2$. This ensures that all the square roots appearing in this formula (on either side, and both the inner ones and the outer ones) have a nonnegative real argument c. Recall here that $\sqrt{c}$ is assigned a unique meaning when c is a nonnegative real number: $\sqrt{c}$ is the unique *non-negative* real number whose square equals c (and so of the two solutions of $x^2 = c$ one *chooses* to denote by $\sqrt{c}$ the non-negative one, for convenience). Under these conditions on a and b one can simply verify the formula by noting (that all the involved square roots are defined in the real numbers, and) that the right-hand side is nonnegative, by squaring both sides and checking that they give the same result after simplification.

PROOF OF THE DOUBLE RADICAL IDENTITY. Writing $R = \sqrt{a^2 - b}$ for brevity, the square of the right-hand side equals

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

Hence the square of the right-hand side of the Double Radical Identity equals the square of the left-hand side of the Double Radical Identity. To conclude a proof of the Double Radical Identity we need to make sure that both sides have the same signs (or are both zero). In fact, our assumptions $a \geq 0$ and $0 \leq b \leq a^2$ imply that all radicals involved are real and non-negative (by convention we choose the non-negative root of a non-negative real number). In particular, the left-hand side is non-negative. Also, of the two radicals at the right-hand side the first is not less than the second (because $a + R \geq a - R$), so their difference is non-negative (in case of the minus sign out of $\pm$). □

This procedure gives a correct and perfectly rigorous *proof* of the identity, but has the drawback that it does not tell us where the identity comes from: we apparently have to recall the identity by heart before we can verify it.

We now show how the identity can be derived from scratch. Hence if we happen to forget it, here is how it can be recovered. As a motivation we may preliminarly note that

if we square a sum $\sqrt{x}+\sqrt{y}$ we get $x+y+2\sqrt{xy}$. Now this latter expression is easily recognisable as the square of $\sqrt{x}+\sqrt{y}$ as long as the terms x and y of the sum are written separately, but not anymore once they are added together. Taking them apart is exactly what our identity aims to achieve. So, given a double radical $\sqrt{a\pm\sqrt{b}}$ (with $0\le b\le a^2$) it is natural to try and express it in the form $\sqrt{a\pm\sqrt{b}}=\sqrt{x}\pm\sqrt{y}$, for some x and y to be determined. Squaring both sides we find $a\pm\sqrt{b}=x+y\pm 2\sqrt{xy}$, at which point we are led to impose

$$\begin{cases} x+y=a \\ 4xy=b. \end{cases}$$

Hence the required x and y are the roots of the quadratic polynomial $z^2-az+b/4$ in the indeterminate z, which of course are $\big(a\pm\sqrt{a^2-b}\big)/2$, leading to the desired identity after choosing $y\le x$.

Example. Let us experiment with the Double Radical Identity on a case where the conditions $a\ge 0$ and $a^2\ge b$ are *not* satisfied. For example, the double radical $\sqrt{2-\sqrt{5}}$ does not represent a real number, because $2-\sqrt{5}<0$. In fact, the condition $a^2\ge b$ is not satisfied here. If we change sign to what is under the square root we get a real radical $\sqrt{-2+\sqrt{5}}$, but the condition $a^2\ge b$ is still not satisfied, and actually $a\ge 0$ is not satisfied either. The Double Radical Identity would give

$$\sqrt{-2+\sqrt{5}}=\sqrt{\frac{-2+\sqrt{(-2)^2-5}}{2}}+\sqrt{\frac{-2-\sqrt{(-2)^2-5}}{2}}=\sqrt{\frac{-2+i}{2}}+\sqrt{\frac{-2-i}{2}},$$

which is not particularly useful as they the simple radicals involve complex numbers. Similarly for the real radical $\sqrt{2+\sqrt{5}}$ the condition $a\ge 0$ is satisfied, but $a\ge 0$ is not, and the Double Radical Identity would give $\sqrt{+2+\sqrt{5}}=\sqrt{1+i/2}+\sqrt{1-i/2}$. The Double Radical Identity may be less useful when complex numbers are involved, but is still correct if properly interpreted. We explore that in the next section.

26. (Optional) Double radicals in the complex case

Now we take a look at the Double Radical Identity more generally, without the above assumptions on a and b. The identity remains valid for a and b any complex numbers, provided we are careful with the meaning of the square roots. In fact, while $\sqrt{a}$ has, by convention, a unique meaning when a is a nonnegative real number, namely, the only positive root of x^2-a, for arbitrary complex $a\neq 0$ the symbol $\sqrt{a}$ actually takes two opposite values, the roots of x^2-a, as it is not possible to make a consistent choice of one root or the other based on algebraic means. In particular, we usually think of $\sqrt{-1}$ as the imaginary unit i, but we could detect no difference in any calculation if we had set

$\sqrt{-1}$ to be $-i$ (as long as we are consistent). [10] When correctly interpreted, the following identity remains valid for any complex numbers a and b:

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

Here any of the six radicals appearing may take two values, which leads to many possible interpretations, only some of which are correct. However, the same radical $\sqrt{a^2-b}$ appears twice on the right-hand side, and whenever that occurs in a formula there is a convention to use the same value for that in both instances (which one being immaterial as it appears with a + sign in one case and a − sign in the other). [11] So in the end the right-hand side generally takes four values, depending on a choice between two values for each of the outside radicals. The left-hand side also generally takes four possible values, depending on a choice between two values for $\sqrt{b}$, and once that choice has been made, on a choice between two values of $\sqrt{a+\sqrt{b}}$. How to match each of the two opposite pairs of values for the right-hand side to each of the two opposite pairs of values for the left-hand side must be made by looking at the arguments of the complex numbers involved, as both possible matches are algebraically equivalent (that is, there is no algebraic way to choose the appropriate match). [12] With this interpretation, the above identity for complex a, b can be simply verified by squaring both sides as we did for the real case.

One application of the double radical identity for complex numbers is to computing square roots of complex numbers. Consider a complex number written in the usual form $a+ib$, where a, b are real numbers. Together with its complex conjugate it can be thought of as $a \pm ib = a + \sqrt{-b^2}$, and the double radical formula gives

$$\begin{aligned}\sqrt{a \pm \sqrt{-b^2}} &= \sqrt{\frac{a+\sqrt{a^2+b^2}}{2}} \pm \sqrt{\frac{a-\sqrt{a^2+b^2}}{2}} \\ &= \sqrt{\frac{\sqrt{a^2+b^2}+a}{2}} \pm i\sqrt{\frac{\sqrt{a^2+b^2}-a}{2}}.\end{aligned}$$

Note that both $\sqrt{a^2+b^2}+a$ and $\sqrt{a^2+b^2}-a$ are nonnegative real numbers, and so both outer radicals in the final expression are square roots of nonnegative real numbers, where we have a convention in place that they usually represent the nonnegative square root. However, because of how we have obtained this expression they should still be considered as complex radicals, each taking two opposite values. Hence the above formula should be

[10] This issue is a little delicate to be discussed further at this point, but it is related with the *conjugation* map $a+ib \mapsto a-ib$ being an *automorphism* of the complex field $\mathbb{C}$.

[11] A similar situation occurs, for example, in Cardano's formulas for the solutions of cubic equations.

[12] In the double radical identity seen at the beginning of the subsection all radicals took nonnegative real values, which amounts to their arguments being 0, rather than the other possible choice π.

interpreted as

$$\pm\sqrt{a \pm \sqrt{-b^2}} = \pm\sqrt{\frac{\sqrt{a^2+b^2}+a}{2}} \pm i\sqrt{\frac{\sqrt{a^2+b^2}-a}{2}},$$

where all the outer radicals are now read as single-valued real radicals. As discussed above, all four $\pm$ signs in the equation are independent (they need not match), and so each side takes four possible values, namely two pairs of opposite values, and the only way to match them correctly is to look at the arguments of the complex numbers involved.

27. Square roots of complex numbers

Solving a quadratic equation with complex coefficients may require taking square roots of complex numbers. Now we look at the problem of computing square roots of complex numbers again, independently of the previous (optional) section on complex double radicals. We will see how computing those square roots in terms of real radicals leads to a biquadratic equation.

Consider a complex number written in the standard form $a + ib$, where $a, b \in \mathbb{R}$, and assume $a + ib \neq 0$ as we may. Of course if the number is written in polar form $a + ib = \rho \cdot (\cos\theta + i\sin\theta)$, where $\rho = \sqrt{a^2+b^2} > 0$ is its modulus and $0 \leq \theta < 2\pi$ is its argument, then we have general formulas for all the nth roots of $a + ib$, namely,

$$\rho^{1/n} \cdot \left(\cos\left(\frac{\theta+2k\pi}{n}\right) + i\sin\left(\frac{\theta+2k\pi}{n}\right)\right), \qquad \text{for } k = 0, \ldots, n-1$$

(or $-n/2 < k \leq n/2$ if we prefer). However, we would like to find the square roots of $a + ib$ algebraically, without using trigonometric functions. More precisely, $a + ib \neq 0$ will have two distinct square roots in the complex numbers, say $x + iy$ and $-x - iy$, with $x, y \in \mathbb{R}$. We would like to compute the real numbers x and y from a and b in an algebraic way, using the four basic operations and possibly radicals.

We have $(x + iy)^2 = a + ib$, that is,

$$x^2 - y^2 + 2ixy = a + ib.$$

Because x, y, a, b are real, this equation is equivalent to the system of equations [13]

$$\begin{cases} x^2 - y^2 = a \\ 2xy = b \end{cases}$$

[13]Note that this is a system of degree $2 \cdot 2 = 4$, and in general we could expect it to have four solutions in the complex numbers. For example, in the special case where $b = 0$ one finds that either $x = 0$ and $y^2 = -a$, or $y = 0$ and $x^2 = a$. However, if $a \neq 0$ then depending on the sign of a exactly one of these two cases will lead to real solutions for both x and y, which are the only acceptable ones for our problem, and produce the two square roots of a in each case.

If we assume $b \neq 0$, which is reasonable because otherwise $a+ib$ is a real number and its square roots are easy to find, then neither x nor y can be zero, and so from the second equation we find $y = b/(2x)$. Substituting this into the first equation we get

$$x^2 - \Big(\frac{b}{2x}\Big)^2 = a,$$

that is,

$$4x^4 - 4ax^2 - b^2 = 0.$$

This is a biquadratic equation, and solving it (without even performing the usual substitution and back, now that we know how it works) we find

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

Because $a^2 < a^2 + b^2$, the right-hand side will only be nonnegative (and actually positive) when we take the + sign in front of the radical, and because our x needs to be real only that case leads to acceptable solutions. Hence $x^2 = \big(a + \sqrt{a^2+b^2}\big)/2$ and, consequently, $y^2 = x^2 - a = \big(-a + \sqrt{a^2+b^2}\big)/2$. In conclusion, the two square roots $\pm(x+iy)$ of $a+ib$ are obtained by taking

$$x = \pm\sqrt{\frac{\sqrt{a^2+b^2}+a}{2}}, \quad \text{and} \quad y = \pm\sqrt{\frac{\sqrt{a^2+b^2}-a}{2}},$$

with matching signs in front of the two main radicals if $b > 0$, and with opposite signs if $b < 0$, as one recognises from the second equation of the system. This can be summarized in the formula

$$\pm\sqrt{a \pm ib} = \pm\sqrt{\frac{\sqrt{a^2+b^2}+a}{2}} \pm i\sqrt{\frac{\sqrt{a^2+b^2}-a}{2}},$$

where the signs have to be appropriately matched as discussed.

Note that there is no need to memorise such complicated formulas: to compute the square roots of a given complex number, just apply the procedure described (see the example below). Nevertheless, let us play with those formulas a bit.

If we write the nonzero complex number $a+ib$ in polar form $a+ib = \rho\cdot(\cos\theta + i\sin\theta)$, where $\rho = \sqrt{a^2+b^2}$ is its modulus and $0 \leq \theta < 2\pi$ is its argument, we can read off the above formulas the halving formulas for sin and cos. In fact, as discussed more generally for nth roots at the beginning of this section, the square roots of $a+ib = \rho(\cos\theta + i\sin\theta)$ will be $\pm\sqrt{\rho}\cdot\big(\cos(\theta/2) + i\sin(\theta/2)\big)$, and so our formulas for the square roots of $a+ib$ imply

$$\cos\Big(\frac{\theta}{2}\Big) = \pm\sqrt{\frac{1+\cos\theta}{2}}, \quad \text{and} \quad \sin\Big(\frac{\theta}{2}\Big) = \pm\sqrt{\frac{1-\cos\theta}{2}},$$

where the signs have to be taken appropriately. Of course these angle halving formulas can also be obtained more directly by inverting the angle duplication formulas

$$\cos\theta = 2\cos^2\Big(\frac{\theta}{2}\Big) - 1 = 1 - 2\sin^2\Big(\frac{\theta}{2}\Big).$$

EXAMPLE. Compute the square roots of the complex number $-3+4i$, expressing them using only square roots of real numbers (that is, expressing their real and complex part using only algebraic operations on real numbers).

We look for real numbers x and y such that $(x+iy)^2 = -3+4i$, that is,

$$x^2 - y^2 + 2ixy = -3 + 4i.$$

Because x, y are real, this equation is equivalent to the system of equations

$$\begin{cases} x^2 - y^2 = -3 \\ 2xy = 4 \end{cases}$$

From the second equation we find $y = 2/x$. Substituting this into the first equation we get $x^2 - (2/x)^2 = -3$, that is, $x^4 + 3x^2 - 4 = 0$, or $(x^2-1)(x^2+4) = 0$. This has roots ± 1 and $\pm 2i$, but because x must be real for our problem we may only take $x = \pm 1$, and correspondingly $y = \pm 2$. In conclusion, the square roots of $-3+4i$ are $1+2i$ and $-1-2i$.

If we had forgotten that x and y are real, and so we accepted $x = \pm 2i$, and correspondingly $y = \mp i$, the argument would not be quite correct but we would still get the correct square roots, as $2i + i(-i) = 1 + 2i$ and its opposite.

REMARK 39. (Optional) It is also possible to compute the square roots of $-3+4i$ by writing it as $-3+\sqrt{-16}$ and applying the Double Radical Identity:

$$\sqrt{-3+4i} = \sqrt{-3+\sqrt{-16}} = \sqrt{\frac{3+\sqrt{3^2+16}}{2}} + \sqrt{\frac{3-\sqrt{3^2+16}}{2}} = \sqrt{1} + \sqrt{-4}.$$

However, as explained in the previous section using the Double Radical Identity leaves sign ambiguities when working with complex numbers, namely, $\sqrt{1}$ can mean 1 or -1, and $\sqrt{-4}$ can mean $2i$ or -21: we now must decide how to correctly match the signs by looking at arguments of the complex numbers involved (or guessing one possible match and square the result to check if it was the correct choice). In essence, the ambiguity which we have is between the square roots of $-3+4i$ and those of $-3-4i$, which are their conjugates.

28. Some special polynomials: self-reciprocal polynomials

The reason why biquadratic polynomials are much easier to factorise than arbitrary quartic polynomials is that they satisfy a special symmetry: they satisfy $f(-x) = f(x)$ (this condition is what, more generally, characterises *even* functions). A similar condition, but involving reciprocals instead of opposites, defines *self-reciprocal* polynomials.

A polynomial $f(x)$ of positive degree n is called *self-reciprocal* if $x^n \cdot f(1/x) = f(x)$. If $f(x) = a_n x^n + \cdots + a_1 x + a_0$ then *its reciprocal polynomial* is

$$x^n \cdot f\left(\frac{1}{x}\right) = a_n + a_{n-1}x + a_{n-2}x^2 + \cdots + a_1 x^{n-1} + a_0 x^n,$$

and so $x^n \cdot f(1/x)$ is a polynomial of degree n, whose coefficients are the coefficients of $f(x)$ read in the opposite order. Hence a polynomial $f(x)$ is self-reciprocal if it equals its reciprocal polynomial, and that means that its sequence of coefficients reads the same backwards as forwards: $a_n = a_0$, $a_{n-1} = a_1$, etc.

The definition of self-reciprocal shows that all roots α are nonzero (because $a_0 = a_n \neq 0$), and that whenever α is a root, its *reciprocal* $1/\alpha$ is a root as well. (This is the reason of the name *self-reciprocal*.) If $f(x)$ is self-reciprocal of odd degree, then one sees at once that -1 is a root, hence $f(x)$ is divisible by $x + 1$, and one can show that the quotient is also self-reciprocal, but of course of even degree. Hence it is enough to see how to deal with a self-reciprocal polynomial of even degree n. We could start with the case of quadratic polynomials, but because we already know how to find their roots in general we pass directly to the case of degree four.

A quartic self-reciprocal polynomial will have the form $ax^4+bx^3+cx^2+bx+a$. The idea for finding its roots, that is, for solving the corresponding equations, is to suitably match x and $1/x$, by taking their sum $x+1/x$. We may start with dividing by x^2 the corresponding equation (as we know that 0 cannot be a root), obtaining $ax^2 + bx + c + b/x + a/x^2 = 0$, that is,

$$a\left(x^2 + \frac{1}{x^2}\right) + b\left(x + \frac{1}{x}\right) + c = 0.$$

Now note that we can express $x^2 + 1/x^2$ in terms of $x + 1/x$,

$$x^2 + \frac{1}{x^2} = \left(x + \frac{1}{x}\right)^2 - 2,$$

and so we can write our equation in the equivalent form

$$a\left(x + \frac{1}{x}\right)^2 + b\left(x + \frac{1}{x}\right) + c - 2a = 0.$$

Now we may set $y = x + 1/x$, and solve the quadratic equation $ay^2 + by + c - 2a = 0$. It may not have any solutions in F, and in that case our quartic self-reciprocal equations cannot have any solutions in F either (because if $\alpha \in F$ were a solution, then $\beta = \alpha + 1/\alpha$ would be a solution of the quadratic equation). If it does have solutions, say β_1 and β_2 (which may be equal), then we may try and solve $x + 1/x = \beta_1$ and $x + 1/x = \beta_2$. Each of these may or may not have solutions in F, giving at most four solutions of our quartic equation. [14]

[14] A self-reciprocal equation of degree six may be dealt with in a similar way, and solving it reduces to solving a cubic equation. In this case we would also have to deal with an expression $x^3 + 1/x^3$, which

EXAMPLE. We use the above method to find all complex roots of the self-reciprocal polynomial $6x^4 + 5x^3 - 38x^2 + 5x + 6$. After equating the polynomial to zero we divide by x^2, rearrange the terms and get

$$6\left(x^2 + \frac{1}{x^2}\right) + 5\left(x + \frac{1}{x}\right) - 38 = 0.$$

Using $(x + 1/x)^2 = x^2 + 2 + 1/x^2$ the equation becomes

$$6\left(x + \frac{1}{x}\right)^2 + 5\left(x + \frac{1}{x}\right) - 50 = 0,$$

which reads $6y^2 + 5y - 50 = 0$ after setting $x + 1/x = y$. This quadratic equation has roots $5/2$ and $-10/3$, and by substituting $y = x + 1/x$ again we find the two equations

$$x + \frac{1}{x} = \frac{5}{2}, \quad \text{and} \quad x + \frac{1}{x} = -\frac{10}{3}.$$

After multiplying by $2x$ or $3x$ they become

$$2x^2 - 5x + 2 = 0, \quad \text{and} \quad 3x^2 + 10x + 3 = 0,$$

whose solutions are 2, $1/2$, and -3, $-1/3$, respectively. Hence these four numbers are the roots of the original polynomial. Because they are all four real, the polynomial factorises into a product of four linear factors already over $\mathbb{R}$, namely,

$$\begin{aligned} 6x^4 + 5x^3 - 38x^2 + 5x + 6 &= 6(x - 2)(x - 1/2)(x + 3)(x + 1/3) \\ &= (x - 2)(2x - 1)(x + 3)(3x + 1). \end{aligned}$$

Because the roots have turned out to be all rational in this example, we could of course also have found them by the Rational Root Test. But we have used a general method for quartic self-reciprocal polynomials, which would work even if there were no rational roots.

29. (Optional) An application: exact trig values of the angle $2\pi/5$

We will compute the exact values of the trigonometric functions (sin and cos, from which the others follows easily) of multiples of the angle $\pi/5$, that is, $36°$. These can be found through geometric arguments. As a general rule, formulas for trigonometric functions are best dealt with by working with the exponential form of complex numbers, and so we set

$$\omega := \exp(2\pi\, i/5) = \cos(2\pi/5) + i\, \sin(2\pi/5).$$

---

can be done by noting that

$$\left(x + \frac{1}{x}\right)^3 = x^3 + 3x + \frac{3}{x} + \frac{1}{x^3} = \left(x^3 + \frac{1}{x^3}\right) + 3\left(x + \frac{1}{x}\right),$$

whence $x^3 + 1/x^3 = (x + 1/x)^3 - 3(x + 1/x) = y^3 - 3y$, etc. There is a formula for solving cubic equations, but we will not see that, and so we stop here with this observation.

Now $\omega^5 = \exp(2\pi\, i/5)^5 = \exp(2\pi\, i) = 1$, and so ω is a fifth root of unity, and hence a root of the polynomial $x^5 - 1$. Because $x^5 - 1 = (x-1)(x^4 + x^3 + x^2 + x + 1)$ and $\omega \neq 1$, we deduce that ω is a root of $x^4 + x^3 + x^2 + x + 1$. Each of ω^2, $\omega^3 = \omega^{-2}$, and $\omega^4 = \omega^{-1}$ is also a fifth root of unity different from 1, hence these numbers are also roots of $x^4 + x^3 + x^2 + x + 1$. Being all distinct, they must be all complex roots of this polynomial, and so

$$x^4 + x^3 + x^2 + x + 1 = (x-\omega)(x-\omega^2)(x-\omega^{-2})(x-\omega^{-1}).$$

Because this polynomial is self-reciprocal polynomial, we have learnt how to compute its roots. After equating the polynomial to zero we divide by x^2, rearrange the terms and get

$$x^2 + \frac{1}{x^2} + x + \frac{1}{x} + 1 = 0.$$

Now because $(x+1/x)^2 = x^2+2+1/x^2$ we can transform the equation into the equivalent equation

$$\left(x + \frac{1}{x}\right)^2 + \left(x + \frac{1}{x}\right) - 1 = 0.$$

After setting $x + 1/x = y$ we solve the resulting equation $y^2 + y - 1 = 0$, and find

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

Now we substitute $y = x + 1/x$ and solve the two equations

$$x + \frac{1}{x} = \frac{-1 \pm \sqrt{5}}{2},$$

which after multiplication by $2x$ become

$$2x^2 - (-1 \pm \sqrt{5})x + 2 = 0.$$

The equation $2x^2 - (-1 + \sqrt{5})x + 2 = 0$ has solutions

$$\frac{\sqrt{5} - 1 \pm \sqrt{(\sqrt{5}-1)^2 - 16}}{4} = \frac{\sqrt{5} - 1 \pm \sqrt{-10 - 2\sqrt{5}}}{4} = \frac{\sqrt{5}-1}{4} \pm i\frac{\sqrt{10+2\sqrt{5}}}{4},$$

and so we conclude that

$$\cos(2\pi/5) = \frac{\sqrt{5}-1}{4}, \quad \text{and} \quad \sin(2\pi/5) = \frac{\sqrt{10+2\sqrt{5}}}{4}.$$

Of course $\sin(2\pi/5)$ could also be computed from $\cos(2\pi/5)$ using the relation $\cos^2\alpha + \sin^2\alpha = 1$. The double radical $\sqrt{10+2\sqrt{5}}$ here cannot be simplified, as $10^2 - (2\sqrt{5})^2 = 80$ is not a perfect square.

Similarly, by solving $2x^2 - (-1 + \sqrt{5})x + 2 = 0$ we find that

$$\cos(4\pi/5) = \frac{-1-\sqrt{5}}{4}, \quad \text{and} \quad \sin(4\pi/5) = \frac{\sqrt{10-2\sqrt{5}}}{4}.$$

From this we obtain

$$\cos(\pi/5) = \frac{1+\sqrt{5}}{4}, \quad \text{and} \quad \sin(\pi/5) = \frac{\sqrt{10-2\sqrt{5}}}{4}.$$

Hence each side of a regular pentagon of radius 1 has length $2\sin(\pi/5) = \big(\sqrt{10-2\sqrt{5}}\big)/2$.

30. Symmetric functions of the roots of quadratic polynomials

In this and the next section we study an important relation between the coefficients of a polynomial and its roots. We start with discussing the case of quadratic polynomials. If the quadratic polynomial ax^2+bx+c (hence with $a \neq 0$) has at least one root in the field F, then we have seen earlier on that it factorises as the product of two polynomials of degree one. By collecting suitable scalar factors those two factors can be taken to have the form $x-\alpha$ and $x-\beta$, and so we have

$$ax^2+bx+c = a(x-\alpha)(x-\beta) = a\big(x^2-(\alpha+\beta)x+\alpha\beta\big).$$

Hence $-b/a = \alpha+\beta$ (the sum of the roots), and $c/a = \alpha\beta$ (the product of the roots). This can be used to guess the roots of a quadratic polynomial in simple cases, but also, more usefully, to use a quadratic equation for solving systems of two equations as in the following example.

Example. Solving the system

$$\begin{cases} x+y = s \\ \quad xy = p \end{cases}$$

in the unknowns x and y, means finding all pairs of numbers x, y whose sum equals s and whose product equals p. One could express y in terms of x using the first equation, hence $y = s-x$, substitute for y in the second equation, solve the corresponding quadratic equation in x obtained, etc., but it is more efficient to exploit the symmetry of the system and to proceed as follows.

The desired x and y will be the roots of the polynomial $(z-x)(z-y)$ in the indeterminate z, which can be written as $z^2-(x+y)z+xy$, and hence equals z^2-sz+p. Therefore, its complex roots are given by the formula $\big(s \pm \sqrt{s^2-4p}\big)/2$. One of the roots will be x, and the other will be y. Of course if the roots are distinct then there are two ways to match x and y to the two roots, and this gives us two solutions (x,y) for our symmetric system. The roots will be equal exactly when $s^2 = 4p$, and in that case the system has a 'double' solution $(x,y) = (s/2, s/2)$.

The polynomials $x+y$ and xy are examples of *symmetric polynomials* in the indeterminates x and y. More generally, a polynomial $f(x,y)$ in x and y is a *symmetric polynomial* if it is unchanged by interchanging the indeterminates x and y, which means

$f(y,x) = f(x,y)$. For example, $x^3 + 2x^2y + 2xy^2 + y^3 + 5xy - 4x - 4y + 7$ is a symmetric polynomial. The special polynomials $x + y$ and xy are called the *elementary symmetric polynomials*.

THEOREM 40. *Every symmetric polynomial $f(x,y)$ (with coefficients in any field F) can be expressed as a polynomial (also with coefficients in F) in the elementary symmetric polynomials $x + y$ and xy.*

This means that if $f(x,y)$ is a symmetric polynomial in x and y (hence with the condition $f(y,x) = f(x,y)$), then $f(x,y) = g(x+y, xy)$, for some polynomial $g(s,p)$ in two indeterminates s and p. More is true: if $f(x,y)$ has integer coefficients, then $g(s,p)$ has integer coefficients as well. We will not prove these statements in general, but just illustrate them with a couple of special cases which can be guessed at once:

$$x^2 + y^2 = (x+y)^2 - 2xy, \qquad x^3 + y^3 = (x+y)^3 - 3xy(x+y).$$

The next case takes a bit more work,

$$\begin{aligned} x^4 + y^4 &= (x+y)^4 - xy(4x^2 + 6xy + 4y^2) \\ &= (x+y)^4 - xy\big(4(x+y)^2 - 2xy\big) \\ &= (x+y)^4 - 4xy(x+y)^2 + 2(xy)^2. \end{aligned}$$

Hence, in the notation of Theorem 40, the symmetric polynomial $f(x,y) = x^4 + y^4$ can be written as $f(x,y) = g(x+y, xy)$, where $g(s,p) = s^4 - 4s^2p + 2p^2$.

More generally, according to Theorem 40 sums $x^n + y^n$ of higher powers can also be expressed as polynomials in $x+y$ and xy (with integer coefficients): one need to do some work as above and use the analogous expressions for smaller values of n. A different and more efficient way of achieving this same goal is described in a later section.

EXAMPLE. The symmetric polynomial $f(x,y) = x^3+2x^2y+2xy^2+y^3+5xy-4x-4y+7$ can be written as

$$f(x,y) = (x+y)^3 - xy(x+y) + 5xy - 4(x+y) + 7 = g(x+y, xy),$$

where $g(s,p) = s^3 - sp - 4s + 5p + 7$.

31. (Optional) The symmetries involved in biquadratic and self-reciprocal polynomials

The methods which we used to solve biquadratic and quartic self-reciprocal equations can be justified in terms of symmetric polynomials.

Indeed, the fact that a biquadratic polynomial $f(x)$ is unchanged when replacing x with $-x$ (that is, $f(-x) = f(x)$, which more generally characterises the *even* polynomials, as opposed to the *odd* polynomials, satisfying $f(-x) = -f(x)$), suggests expressing it in

terms of the 'elementary symmetric polynomials' in x and $-x$, which are $x+(-x)=0$ and $x\cdot(-x)=-x^2$. (We have slightly abused language here, as x and $-x$ are not independent indeterminates.) This is, in fact, what we did, thinking of a biquadratic polynomial as a polynomial in x^2 (which makes no practical difference from using $-x^2$ instead).

Similarly, a self-reciprocal polynomial $f(x)$ of degree n satisfies $x^n \cdot f(1/x) = f(x)$, hence it is not quite left unchanged by replacing x with $1/x$, but almost. In fact, our first step in finding the roots of $f(x)$ (and then factorising it), for even n, was dividing $f(x)$ by $x^{n/2}$. Now the rational expression (that is, quotient of two polynomials) $g(x) = f(x)/x^{n/2}$ satisfies $g(1/x) = g(x)$, because

$$g(1/x) = \frac{f(1/x)}{(1/x)^{n/2}} = x^{n/2} \cdot f(1/x) = f(x)/x^{n/2} = g(x).$$

Hence $g(x)$ is left unchanged by replacing x with $1/x$, and as a consequence of Theorem 40 (which extends to quotients of polynomials) it can be expressed in terms of the 'elementary symmetric polynomials' in x and $1/x$, which are $x+1/x$ and $x\cdot 1/x = 1$. This is precisely what we did when finding the roots of self-reciprocal polynomials: we expressed x^2+1/x^2 as a polynomial in $x+1/x$ (and we would do the same with each x^k+1/x^k if we wished to deal with self-reciprocal polynomials of even degree higher than 4).