STRAIGHTEDGE & COMPASS *VERSUS* ORIGAMI

NOTES FOR THE "REAL MATHS": THE CHRISTMAS EDITION

JIŘÍ VELEBIL

> My origami creations, in accordance with the laws of nature, require the use of geometry, science, and physics. They also encompass religion, philosophy, and biochemistry. Overall, I want you to discover the joy of creation by your own hand. The possibility of creation from paper is infinite.
>
> *Akira Yoshizawa*, Japanese Origami Grand Master, 14 March 1911 – 14 March 2005

A brief synopsis. Chapter 2 is devoted to (mainly) the recollection of known facts about *fields* and *vector spaces*. We apply this theory to develop the very basics of the topic called *field extensions*. Chapter 3 deals with a general procedure of producing field extensions by means of *polynomials*. In Chapter 4 we state (without proof) the celebrated *Hermite-Lindemann-Weierstrass Theorem* that shows that certain real numbers cannot be roots of polynomials with rational coefficients. In particular, π is such a real number — a result that we will need in discussing *the problem of squaring the circle*.

In Chapter 5 we show that the collection of lengths that one can construct by straightedge and compass forms a *positive cone* $\mathbb{E}_+$ *of an ordered field* $\mathbb{E}$. The resulting field $\mathbb{E}$ is called the *field of constructible numbers* or the *field of Euclidean numbers*. We give a characterisation of elements of $\mathbb{E}$.

The characterisation of $\mathbb{E}$ allows one to prove that certain well-known problems cannot have a solution when working with compass and straightedge only: *the trisection of an angle*, *the squaring of the circle*, *the doubling of the cube*. This is the contents of Chapter 6.

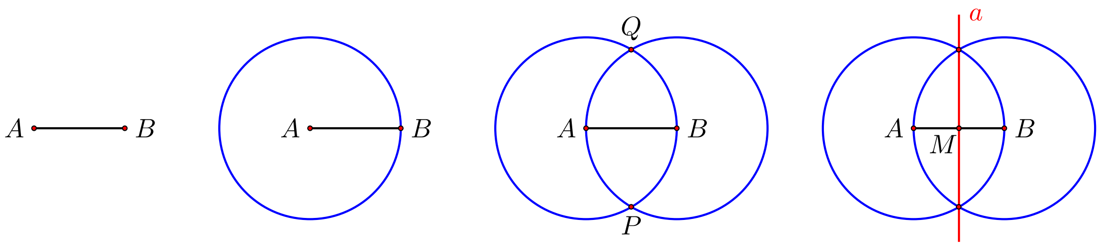
In Chapter 7 we introduce the *Huzita-Hatori axiomatics* for paper folding. These axioms are motivated by the study of *origami*. We then show in Chapter 8 that Huzita-Hatori axioms are stronger than constructions with compass and straightedge. In fact, there exists a *field* $\mathbb{O}$ *of origami numbers* that extends the field $\mathbb{E}$. As a result, some classical problems of Ancient Greek geometry can be solved by the origami constructions. Alas, *squaring the circle cannot be solved even by paper folding*.

In final Chapter 9 we comment a bit on *Galois Theory* — the branch of mathematics that was inspired by the compass and straightedge constructions.

1. Intro — Constructions by straightedge & compass

Geometry achieved one of its peaks undoubtedly in Ancient Greece. In fact, the work in geometry, presented in *Elements* by Euclid [11] roughly 300 BC, is a standard of mathematical reasoning till today. From Euclid's work it follows that very many problems in geometry can be solved by using an *unmarked straightedge* and a *compass*. We present one problem and its well-known solution in the next example.

1.1. **Example (Construction of a midpoint of a segment by straightedge and compass).** Given a line segment AB in the plane, can one construct its *midpoint*, i.e., can one construct a point M on the segment AB, such that the segments AM and MB have the same length?



Step 1 Step 2 Step 3 Step 4

There is a well-known construction that involves the four steps drawn above:

Date: 6 December 2019.

(Step 1): Draw the segment AB using straightedge.
(Step 2): Using compass, draw the circle with centre at A and radius the length of the segment AB.
(Step 3): Using compass, draw the circle with centre at B and radius the length of the segment AB.
(Step 4): Mark the intersection of the two circles by P and Q and, using, straightedge, draw the line $a = PQ$. The midpoint M is the intersection of a and AB.

How do we know that we have indeed constructed the midpoint of AB? Well, one has to argue about the *congruence* of the four triangles $\triangle AMQ$, $\triangle BMQ$, $\triangle BMP$ and $\triangle BMQ$ in the picture. This is left as an easy exercise.

It is worth noting that the proof that one can find the midpoint M of a segment AB is given in Proposition 10 of Book I of [11]. Euclid proves his proposition *differently*, however. We chose to give a "more modern proof" by *Apollonius of Perga*.[1]

Example 1.1 shows all the important aspects of constructions with unmarked straightedge and compass:

(1) We can draw a segment of a line, connecting two points.
(2) We can draw a line through two distinct points.
(3) We can draw a circle with a centre in a given point and with a given radius.
(4) As *constructible*, we regard only points arising as intersections of lines and circles drawn by points (1)–(3) above.

In what follows, we will make all of the above precise and we will characterise constructible points.

The problem is best tackled by the technique of *adding elements to a given field*. That is why we proceed by introducing some more advanced concepts of field theory.

2. Fields and their extensions

Recall from, e.g., [43] that a field is an algebraic structure that generalises the properties of addition and multiplication of real numbers. Although in this text we will deal mainly with *subfields* of the field $\mathbb{R}$ of reals, we will recall the general definition of a *field extension* here. In particular, we will be interested in field extensions that are determined by quadratic equations.

2.1. **Definition (Commutative rings with a unit and fields).** A *commutative ring with a unit* is given by a set $\mathbb{F}$ (a typical member of $\mathbb{F}$ will be denoted by r and thought of as a "number") that is equipped with two functions

$$+ : \mathbb{F} \times \mathbb{F} \longrightarrow \mathbb{F} \quad \text{(read: \textit{the addition})}, \qquad \cdot : \mathbb{F} \times \mathbb{F} \longrightarrow \mathbb{F} \quad \text{(read: \textit{the multiplication})},$$

that are subject to the following three sets of axioms:

(1) Axioms for addition.
  (a) The existence of zero.
  There is an element 0 in $\mathbb{F}$ such that, for all r in $\mathbb{F}$, such that $r + 0 = 0 + r = r$ holds for all r in $\mathbb{F}$. The element 0 is called *zero*.
  (b) The commutativity of addition.
  For all r, s in $\mathbb{F}$, the equality $r + s = s + r$ holds.
  (c) The associativity of addition.
  For all r, s, t in $\mathbb{F}$, the equality $r + (s + t) = (r + s) + t$ holds.
  (d) The existence of an additive inverse.
  For every r in $\mathbb{F}$ there is a unique s such that the equality $r + s = s + r = 0$ holds. The uniquely determined s is called the *additive inverse* to r and it is denoted by $-r$.
(2) Axioms for multiplication.
  (a) The existence of unit.
  There is an element 1 such that the equalities $1 \cdot r = r = r \cdot 1$ holds for all r in $\mathbb{F}$. The element 1 is called *unit*.
  (b) The commutativity of multiplication.
  For all r, s in $\mathbb{F}$, the equality $r \cdot s = s \cdot r$ holds.

[1] Apollonius of Perga (cca 300BC–200BC) was an outstanding Greek geometer. Based on the work of Euclid and Archimedes, he worked mainly on conic sections. He gave definitions of the terms *ellipse*, *parabola*, and *hyperbola* that we still use today. See Remark 7.3 below.

(c) The associativity of multiplication.
For all r, s, t in $\mathbb{F}$ the equality $r \cdot (s \cdot t) = (r \cdot s) \cdot t$ holds.

(3) The distributive law.
For all r, s, t the equality $r \cdot (s + t) = (r \cdot s) + (r \cdot t)$ holds.

If, moreover, the following axiom

(4) The invertibility test.
For every r in $\mathbb{F}$, the inequality $r \neq 0$ holds if and only if there is a unique r^{-1} such that the equalities $r \cdot r^{-1} = 1 = r^{-1} \cdot r$ holds. The uniquely determined r^{-1} is called the (multiplicative) *inverse* of r.

holds, then $\mathbb{F}$ is called a *field*.

2.2. **Definition (Ordered field).** A field $\mathbb{F}$ is called *ordered*, if there is a subset $\mathbb{F}_+$ in $\mathbb{F}$ (the so-called *positive cone of* $\mathbb{F}$) with the following properties:

(1) For every r in $\mathbb{F}$ we have either $-r \in \mathbb{F}_+$, or $r = 0$, or $r \in \mathbb{F}_+$.
(2) If $r \in \mathbb{F}_+$ and $s \in \mathbb{F}_+$, then $r + s \in \mathbb{F}_+$ and $rs \in \mathbb{F}_+$ hold.

In every ordered field we denote by

$$r < s$$

the fact that $s - r \in \mathbb{F}_+$.

We will also use the notation

$$r \leq s$$

to denote that $r < s$ or $r = s$ holds.

We will mostly work with *ordered* fields in what follows. Recall (see, e.g., [43]) that in an ordered field $\mathbb{F}$ one has

$$r^2 \geq 0 \quad \text{for all } r \text{ in } \mathbb{F} \qquad \text{and} \qquad r^2 = 0 \quad \text{iff} \quad r = 0$$

In particular, the inequalities

$$1 > 0 \qquad \text{and} \qquad 2 = 1 + 1 > 0$$

hold in any ordered field.

2.3. **Task for you.** *Suppose $k = r_1^2 = r_2^2$ holds for $r_1 \geq 0$ and $r_2 \geq 0$ in an ordered field $\mathbb{F}$. Then $r_1 = r_2$.*

Hints for the proof. Clearly, we are dealing with $k \geq 0$. The assertion is clear if $k = 0$. Suppose therefore that $k > 0$. Then $r_1 > 0$ and $r_2 > 0$ must hold. Moreover, the equalities

$$0 = r_1^2 - r_2^2 = (r_1 - r_2) \cdot (r_1 + r_2)$$

hold. If $r_1 + r_2 = 0$ holds, then $r_1 = -r_2$, which contradicts the assumption that $r_1 > 0$, $r_2 > 0$. Therefore $r_1 - r_2 = 0$ must hold. Equivalently, $r_1 = r_2$ must hold, as desired. ■

The above properties of squaring in ordered fields allow us to define the first important concept: the *square root* of a non-negative element.

2.4. **Definition (Square roots in an ordered field).** We say that an element k of an ordered field $\mathbb{F}$ has a *square root* in $\mathbb{F}$, if there is $r \geq 0$ in $\mathbb{F}$ such that $r^2 = k$ holds.[a] We denote the square root r of k by

$$\sqrt{k}$$

[a] Observe that no $k < 0$ can have a square root in an ordered field.

The proof of Task 2.3 suggests that we should better understand solutions of quadratic equations in a general ordered field. As the next example shows, solving quadratic equations in an ordered field is not different from the familiar procedure of solving quadratic equations in the ordered field $\mathbb{R}$ of reals.

2.5. **Example** (**Quadratic equations in an ordered field**). In what follows we will use the knowledge of the form of a solution of *quadratic equations*

$$ax^2 + bx + c = 0$$

where a, b, c are in $\mathbb{F}$ and $a \neq 0$. We derive now the well-known formula for the solution in case the field $\mathbb{F}$ is *ordered*. Since $a \neq 0$, we can as well derive a formula for solving the equation

$$x^2 + b \cdot a^{-1} \cdot x + c \cdot a^{-1} = 0$$

and, by putting

$$B = b \cdot a^{-1} \qquad C = c \cdot a^{-1}$$

it suffices to derive a formula for solving the equation

$$x^2 + Bx + C = 0$$

where B and C are in $\mathbb{F}$. To that end, express the left-hand side as follows:

$$\begin{aligned} x^2 + Bx + C &= x^2 + 2 \cdot (B \cdot 2^{-1}) \cdot x + (B \cdot 2^{-1})^2 - (B \cdot 2^{-1})^2 + C \\ &= \Big(x + B \cdot 2^{-1}\Big)^2 - (B \cdot 2^{-1})^2 + C \end{aligned}$$

Above, we have used the fact that $2 \neq 0$ holds in any ordered field. Thus, we are to solve the equation

$$\Big(x + B \cdot 2^{-1}\Big)^2 - (B \cdot 2^{-1})^2 + C = 0$$

or, equivalently, the equation

$$\Big(x + B \cdot 2^{-1}\Big)^2 = \underbrace{(B \cdot 2^{-1})^2 - C}_{=\Delta}$$

The expression $\Delta = (B \cdot 2^{-1})^2 - C$ is called the *discriminant* of the equation $x^2 + Bx + C = 0$, since it discriminates, whether a solution exists.

(1) Suppose $\Delta = 0$. Then we are to solve the equation

$$\Big(x + B \cdot 2^{-1}\Big)^2 = 0$$

This equation is clearly satisfied if and only if $x + B \cdot 2^{-1} = 0$, or, equivalently, if and only if

$$x = -B \cdot 2^{-1}$$

(2) Suppose $\Delta < 0$. Since $\mathbb{F}$ is an ordered field, $r^2 \geq 0$ for any $r \in \mathbb{F}$. Hence the equality

$$\Big(x + B \cdot 2^{-1}\Big)^2 = \underbrace{(B \cdot 2^{-1})^2 - C}_{=\Delta<0}$$

has no solution in $\mathbb{F}$.

(3) Suppose $\Delta > 0$. Then one of the following two cases can happen:

(a) $\Delta = r^2$ for some r in $\mathbb{F}$. Then the square root

$$\sqrt{\Delta}$$

of Δ exists in $\mathbb{F}$. Then the equation

$$\Big(x + B \cdot 2^{-1}\Big)^2 = \underbrace{(B \cdot 2^{-1})^2 - C}_{=\Delta=r^2}$$

can be rewritten as

$$\Big(x + B \cdot 2^{-1}\Big)^2 - k^2 = (x + B \cdot 2^{-1} + k) \cdot (x + B \cdot 2^{-1} - k) = 0$$

Hence, in this case, we have

$$x = -B \cdot 2^{-1} - k = -B \cdot 2^{-1} - \sqrt{\Delta} \quad \text{or} \quad x = -B \cdot 2^{-1} + k = -B \cdot 2^{-1} + \sqrt{\Delta}$$

which we write as

$$x_{1,2} = -B \cdot 2^{-1} \pm \sqrt{\Delta}$$

(b) There is no r in $\mathbb{F}$ such that $\Delta = r^2$ holds. In this case the square root of Δ does not exist in $\mathbb{F}$. Clearly, there can be no solution of

$$\Big(x + B \cdot 2^{-1}\Big)^2 = \underbrace{(B \cdot 2^{-1})^2 - C}_{=\Delta \neq r^2 \text{ for all } r \text{ in } \mathbb{F}}$$

in $\mathbb{F}$ in this case.

To summarise: the equation $x^2 + Bx + C = 0$ either has a solution of the form

$$x_{1,2} = -B \cdot 2^{-1} \pm \sqrt{\Delta} \qquad \text{in case when } \Delta = (B \cdot 2^{-1})^2 - C \text{ has a square root in } \mathbb{F}$$

or no solution otherwise.

Coming back to the substitution

$$B = b \cdot a^{-1} \qquad C = c \cdot a^{-1}$$

it is now an easy exercise to show that the equation $ax^2 + bx + c = 0$ either

(1) has a solution of the form

$$x_{1,2} = (2 \cdot a)^{-1} \cdot (-b \pm \sqrt{b^2 - 4ac}) \qquad \text{in case when } b^2 - 4ac \text{ has a square root in } \mathbb{F}$$

or

(2) has no solution.

Thus, the usual formulas for solving quadratic equations hold in every ordered field $\mathbb{F}$.

By the above, not every quadratic equation needs to have a solution in an ordered field. We will be particularly interested in quadratic equations that cannot be solved, and we give the name to the respective quadratic polynomials.

2.6. **Definition (Irreducible quadratic polynomials).** Let $\mathbb{F}$ be an ordered field. A polynomial $ax^2 + bx + c$ with a, b, c in $\mathbb{F}$, $a \neq 0$ is called *irreducible*, if the quadratic equation $ax^2 + bx + c = 0$ has no solution in $\mathbb{F}$.

2.7. **Remark (General irreducible polynomials).** In Definition 3.4 below we give the definition of an *irreducible polynomial of a general degree.*

2.8. **Definition ((Ordered) subfields).** Given two fields $\mathbb{F}_1$ and $\mathbb{F}_2$, we call $\mathbb{F}_1$ to be a *subfield* of $\mathbb{F}_2$ (or, equivalently, that $\mathbb{F}_2$ is an *extension* of $\mathbb{F}_1$), if the following properties hold:

(1) $\mathbb{F}_1$ is a subset of $\mathbb{F}_2$.

(2) The set $\mathbb{F}_1$ is closed in $\mathbb{F}_2$ under addition, multiplication, zero, unit and taking inverses.

If the fields $\mathbb{F}_1$ and $\mathbb{F}_2$ are ordered fields, we say that $\mathbb{F}_1$ is an *ordered subfield* of $\mathbb{F}_2$ (or, equivalently, that $\mathbb{F}_2$ is an *ordered extension* of $\mathbb{F}_1$), if $\mathbb{F}_1$ is a subfield of $\mathbb{F}_2$ and, moreover, $(\mathbb{F}_1)_+ \subseteq (\mathbb{F}_2)_+$ holds.

If $\mathbb{F}_2$ extends $\mathbb{F}_1$, we denote this fact by[a]

$$\mathbb{F}_1 \preceq \mathbb{F}_2$$

[a] The notation $\preceq$ is not standard, however. In literature one often finds $\mathbb{F}_2/\mathbb{F}_1$ what we denote by $\mathbb{F}_1 \preceq \mathbb{F}_2$. We chose a different notation in order not to confuse the reader with the notation $\mathbb{F}[x]/p(x)$ that we use for quotient rings of rings of polynomials, see Task 3.2 below.

In what follows we will need a tiny bit of linear algebra. We recall the well-known definition of a vector space, for convenience.

2.9. **Definition (Vector spaces over a field).** A *vector space over a field* $\mathbb{F}$ is a set V, whose elements are called *vectors* (and they will be denoted by $\vec{x}$, etc.), together with two operations

$+ : V \times V \longrightarrow V$ (read: *the addition of vectors*), $\cdot : \mathbb{F} \times V \longrightarrow V$ (read: *the multiplication by scalars*),

that are subject to the following four axioms:

(1) Axioms for addition of vectors.

(a) The existence of zero vector.
There is an element $\vec{o}$ in V such that, for all $\vec{x}$ in V, the equality $\vec{x} + \vec{o} = \vec{o} + \vec{x} = \vec{x}$ holds. The element $\vec{o}$ is called *zero vector*.

(b) The commutativity of addition.
For all $\vec{x}$, $\vec{x'}$ in V, the equality $\vec{x} + \vec{x'} = \vec{x'} + \vec{x}$ holds.

(c) The associativity of addition.
For all $\vec{x}$, $\vec{x'}$, $\vec{x''}$ in V, the equality $\vec{x} + (\vec{x'} + \vec{x''}) = (\vec{x} + \vec{x'}) + \vec{x''}$ holds.

(d) The existence of an additive inverse.
For every $\vec{x}$ in V there is a unique $\vec{y}$ in V such that the equality $\vec{x} + \vec{y} = \vec{y} + \vec{x} = \vec{o}$ holds. The uniquely determined s is called the *additive inverse* to $\vec{x}$ and it is denoted by $-\vec{x}$.

(2) Axioms for multiplication by scalars.

(a) Multiplication by unit.
The equality $1 \cdot \vec{x} = \vec{x}$ holds for all $\vec{x}$ in V.

(b) The associativity of multiplication.
For all r, s in $\mathbb{F}$ and all $\vec{x}$ in V the equality $r \cdot (s \cdot \vec{x}) = (r \cdot s) \cdot \vec{x}$ holds.

(3) The distributive laws.
For all r, s in $\mathbb{F}$ and all $\vec{x}$, $\vec{x'}$ in V the equalities $r \cdot (\vec{x} + \vec{x'}) = (r \cdot \vec{x}) + (r \cdot \vec{x'})$ and $(r+s) \cdot \vec{x} = (r \cdot \vec{x}) + (s \cdot \vec{x'})$ hold.

2.10. **Task for you (Field extensions are vector spaces over subfields).** *Suppose $\mathbb{F}_1$ and $\mathbb{F}_2$ are fields, and suppose $\mathbb{F}_1 \preceq \mathbb{F}_2$. Then $\mathbb{F}_2$ is a vector space over the field $\mathbb{F}_1$ with the operations defined as expected:*

(1) *"Vector addition": the addition of x, x' in $\mathbb{F}_2$ as vectors is $x \underset{\mathbb{F}_2}{+} x'$.*

(2) *"Scalar multiplication": the multiplication of a vector x in $\mathbb{F}_2$ by a scalar r in $\mathbb{F}_1$ is $r \underset{\mathbb{F}_2}{\cdot} x$.*

Hints for the proof. The proof is really straightforward: go through the axioms for vector spaces in Definition 2.9. ■

2.11. **Definition (Degree of an extension).** Suppose $\mathbb{F}_1$ and $\mathbb{F}_2$ are fields, and suppose $\mathbb{F}_1 \preceq \mathbb{F}_2$. The dimension of $\mathbb{F}_2$ as a vector space over $\mathbb{F}_1$ is called the *degree* of $\mathbb{F}_2$ over $\mathbb{F}_1$ and denoted by $[\mathbb{F}_2 : \mathbb{F}_1]$.

2.12. **Task for you (The tower of extensions).** *Suppose that $\mathbb{F}_1 \preceq \mathbb{F}_2 \preceq \mathbb{F}_3$ holds, with $[\mathbb{F}_2 : \mathbb{F}_1]$ and $[\mathbb{F}_3 : \mathbb{F}_2]$ being finite. Then $[\mathbb{F}_3 : \mathbb{F}_1]$ is finite and the equality*

$$[\mathbb{F}_3 : \mathbb{F}_1] = [\mathbb{F}_3 : \mathbb{F}_2] \cdot [\mathbb{F}_2 : \mathbb{F}_1]$$

holds.

Hints for the proof. Let $[\mathbb{F}_2 : \mathbb{F}_1] = a$ and $[\mathbb{F}_3 : \mathbb{F}_2] = b$ and let $\{\alpha_1, \ldots, \alpha_a\}$ be the basis of $\mathbb{F}_2$ over $\mathbb{F}_1$ and $\{\beta_1, \ldots, \beta_b\}$ be the basis of $\mathbb{F}_3$ over $\mathbb{F}_2$.

We will prove that the set

$$B = \{\alpha_i \cdot \beta_j \mid i \in a,\ j \in b\}$$

of ab-many elements of $\mathbb{F}_3$ forms a basis of $\mathbb{F}_3$ over $\mathbb{F}_1$.

(1) B generates $\mathbb{F}_3$ over $\mathbb{F}_1$.
Let r be any element of $\mathbb{F}_3$. Since $\{\beta_1, \ldots, \beta_b\}$ is the basis of $\mathbb{F}_3$ over $\mathbb{F}_2$, we can write

$$r = \sum_{j=1}^{b} r_j \cdot \beta_j$$

for some r_j in $\mathbb{F}_2$. Since $\{\alpha_1, \dots, \alpha_a\}$ is the basis of $\mathbb{F}_2$ over $\mathbb{F}_1$, we can write for every $j \in b$

$$r_j = \sum_{i=1}^{a} r_{ij} \cdot \alpha_i$$

for some r_{ij} in $\mathbb{F}_1$.

Therefore

$$r = \sum_{j=1}^{b} \Big(\sum_{i=1}^{a} r_{ij} \cdot \alpha_i \Big) \cdot \beta_j = \sum_{j=1}^{b} \sum_{i=1}^{a} r_{ij} \cdot \alpha_i \cdot \beta_j$$

Thus, B generates $\mathbb{F}_3$ over $\mathbb{F}_1$.

(2) B is linearly independent over $\mathbb{F}_1$.

Suppose

$$\sum_{j=1}^{b} \sum_{i=1}^{a} a_{ij} \cdot \alpha_i \cdot \beta_j = 0$$

for some a_{ij} in $\mathbb{F}_1$. Since

$$\sum_{j=1}^{b} \sum_{i=1}^{a} a_{ij} \cdot \alpha_i \cdot \beta_j = \sum_{j=1}^{b} \Big(\sum_{i=1}^{a} a_{ij} \cdot \alpha_i \Big) \cdot \beta_j$$

and since $\{\beta_1, \dots, \beta_b\}$ is linearly independent, we have

$$\sum_{i=1}^{a} a_{ij} \cdot \alpha_i = 0$$

for all $j \in b$. Since $\{\alpha_1, \dots, \alpha_a\}$ is linearly independent, we have

$$a_{ij} = 0$$

for all $i \in a$ and all $j \in b$.

The proof is finished. ■

Recall that vector spaces can also have an infinite dimension. We will deal with vector spaces of *finite* dimensions only. In particular, we will first study fields $\mathbb{F}_2$ that are vector spaces of dimension 2 over $\mathbb{F}_1$. Such extensions are called *quadratic*. A full characterisation of quadratic extensions of ordered subfields of $\mathbb{R}$ is given in Task 2.15. We first show a typical quadratic extension of an ordered field.

2.13. **Task for you (Quadratic extension of an ordered field).** *Suppose $\mathbb{F}$ is any ordered field. Let k be an element of $\mathbb{F}$ that does not have a square root in $\mathbb{F}$. Then the set*

$$\mathbb{F}(\sqrt{k}) = \{a + b\sqrt{k} \mid a, b \in \mathbb{F}\}$$

together with the operations

$$\begin{aligned} (a + b\sqrt{k}) + (a' + b'\sqrt{k}) &= (a + a') + (b + b')\sqrt{k} \\ (a + b\sqrt{k}) \cdot (a' + b'\sqrt{k}) &= (aa' + bb'k) + (ab' + a'b)\sqrt{k} \end{aligned}$$

is a field again. Moreover, $\mathbb{F} \preceq \mathbb{F}(\sqrt{k})$ holds and $[\mathbb{F}(\sqrt{k}) : \mathbb{F}] = 2$. [a]

[a] The field $\mathbb{F}(\sqrt{k})$ is called the *quadratic extension* of $\mathbb{F}$. Notice that $\sqrt{k}$ is just an *abstract symbol*, since the square root of k in $\mathbb{F}$ is assumed not to exist in $\mathbb{F}$.

HINTS FOR THE PROOF. Let us notice first the following facts:

(1) If we identify $a + 0 \cdot \sqrt{k}$ with a, we can regard $\mathbb{F}$ as a subset of $\mathbb{F}(\sqrt{k})$. Moreover, $\mathbb{F} \preceq \mathbb{F}(\sqrt{k})$ would hold under this identification, as soon as we proved that $\mathbb{F}(\sqrt{k})$ is a field.

(2) The equality

$$(a+b\sqrt{k})\cdot(a-b\sqrt{k}) = a^2 - b^2k$$

holds in $\mathbb{F}(\sqrt{k})$. Moreover,

$$a^2 - b^2k = \begin{cases} b^2\cdot\underbrace{\Big((a\cdot b^{-1})^2 - k\Big)}_{\neq 0} \neq 0, & \text{if } b\neq 0 \\ a^2, & \text{if } b = 0\end{cases}$$

Indeed, use the definition of multiplication in $\mathbb{F}(\sqrt{k})$ and our assumption on k.

Verifying the field axioms for $\mathbb{F}(\sqrt{k})$ is easy, save perhaps for the existence of inverses. If $a+b\sqrt{k}\neq 0$ in $\mathbb{F}(\sqrt{k})$, then it cannot be the case that both $a=0$ and $b=0$. By (2) above, $a^2-b^2k\neq 0$. Then the inverse of $a+b\sqrt{k}$ is given by the formula

$$(a+b\sqrt{k})^{-1} = a\cdot(a^2-b^2k)^{-1} - b\cdot(a^2-b^2k)^{-1}\cdot\sqrt{k}$$

Since $\{1,\sqrt{k}\}$ is clearly a basis of $\mathbb{F}(\sqrt{k})$ over $\mathbb{F}$, the equality $[\mathbb{F}(\sqrt{k}):\mathbb{F}]=2$ holds. ■

2.14. **Task for you (Adjoining elements to a field).** *Suppose $\mathbb{F}_1 \preceq \mathbb{F}_2$ and let α_1, …, α_k be in $\mathbb{F}_2$. Then there exists a smallest field*[a]

$$\mathbb{F}_1(\alpha_1,\dots,\alpha_k)$$

such that

$$\mathbb{F}_1 \preceq \mathbb{F}_1(\alpha_1,\dots,\alpha_k) \preceq \mathbb{F}_2$$

holds and such that $\mathbb{F}_1(\alpha_1,\dots,\alpha_k)$ contains all α_1, …, α_k.

[a] The field $\mathbb{F}_1(\alpha_1,\dots,\alpha_k)$ is called the *extension of $\mathbb{F}_1$ by α_1, …, α_k.*

Hints for the proof. Consider the set

$$\mathscr{F} = \{\mathbb{F} \mid \mathbb{F} \text{ is a field containing all } \alpha_1,\ \dots,\ \alpha_k, \text{ and } \mathbb{F}_1\preceq\mathbb{F}\preceq\mathbb{F}_2 \text{ holds}\}$$

Then $\mathscr{F}\neq\emptyset$, since $\mathbb{F}_2\in\mathscr{F}$ holds. Define

$$\mathbb{F}_1(\alpha_1,\dots,\alpha_k) = \bigcap\mathscr{F}$$

It is clear that $\mathbb{F}_1(\alpha_1,\dots,\alpha_k)$ is a field and that it has all the desired properties. ■

2.15. **Task for you (Characterisation of quadratic extensions of ordered subfields of $\mathbb{R}$).** *Suppose that $\mathbb{F}_1$, $\mathbb{F}_2$ are ordered fields, and let $\mathbb{F}_1\preceq\mathbb{F}_2\preceq\mathbb{R}$. Then the following are equivalent:*

1. *$[\mathbb{F}_2:\mathbb{F}_1]=2$, i.e., $\mathbb{F}_2$ is a quadratic extension of $\mathbb{F}_1$.*
2. *$\mathbb{F}_2=\mathbb{F}_1(\alpha)$, where $\alpha\in\mathbb{R}$ solves a quadratic equation $ax^2+bx+c=0$ for an irreducible polynomial ax^2+bx+c with coefficients in $\mathbb{F}_1$.*
3. *$\mathbb{F}_2=\mathbb{F}_1(\sqrt{k})$ for some k in $\mathbb{R}$ that does not have a square root in $\mathbb{F}_1$.*

Hints for the proof. (1) implies (2). Let α be any element of $\mathbb{F}_2$, which is not in $\mathbb{F}_1$. Then $\{1,\alpha,\alpha^2\}$ is linearly dependent, i.e., there are a, b, c in $\mathbb{F}_1$ such that $a\alpha^2+b\alpha+c=0$. Since α is not in $\mathbb{F}_1$, we have $a\neq 0$. Thus, α is the root of the quadratic polynomial ax^2+bx+c. The polynomial is clearly irreducible.

(2) implies (3). Since the roots α_1, α_2 of $ax^2+bx+c=0$ are given by the formula

$$\alpha_{1,2} = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

see Example 2.5, the assertion follows.

(3) implies (1). Clearly, $\{1, \sqrt{k}\}$ is a basis of $\mathbb{F}_2$ over $\mathbb{F}_1$. See Task 2.13. ■

2.16. **Definition (Polynomials over a field).** Let $\mathbb{F}$ be a field. A *polynomial in indeterminate x with coefficients in $\mathbb{F}$* is an expression

$$\text{either} \quad 0 \text{ (a polynomial of degree } -\infty) \qquad \text{or} \quad \sum_{i=1}^{n} a_k x^k \text{ (a polynomial of degree } n)$$

where n is a positive natural number, $a_0, \ldots, a_n$ are in $\mathbb{F}$ and $a_n \neq 0$.

2.17. **Remark (Polynomials are not the same thing as functions).** Given a polynomial $p(x)$ in $\mathbb{F}[x]$, it is tempting to substitute various $a \in \mathbb{F}$ for x and treat thus $p(x)$ as a *function* p from $\mathbb{F}$ to $\mathbb{F}$. For example, given a polynomial $p(x) = -3x^2 + 5x + 7$ in $\mathbb{R}[x]$, we have $p(-2) = -3 \cdot (-2)^2 + 5 \cdot (-2) + 7 = -15$.

This approach involves a tiny problem, though. Consider the following *different* polynomials $p(x)$, $q(x)$ over the field $\mathbb{Z}_2$:

$$p(x) = x + 1 \quad \text{and} \quad q(x) = x^2 + 1$$

We clearly have $p(x) \neq q(x)$ as *expressions*, but the equalities

$$\begin{array}{ll} p(0) = 1 & \qquad q(0) = 1 \\ p(1) = 0 & \qquad q(1) = 0 \end{array}$$

show that $p = q$ as *functions*.

Such an example can be found whenever $\mathbb{F}$ is a *finite* field. Indeed, the set $\mathbb{F}[x]$ of all polynomials with coefficients in $\mathbb{F}$ is easily shown to be infinite, whereas the set of all maps from $\mathbb{F}$ to $\mathbb{F}$ is finite.

2.18. **Definition (Addition and multiplication of polynomials).** Suppose $a(x)$ and $b(x)$ are polynomials in $\mathbb{F}[x]$. Define their *addition* as the polynomial

$$a(x) + b(x) = \begin{cases} a(x), & \text{if } b(x) = 0 \\ b(x), & \text{if } a(x) = 0 \\ \displaystyle\sum_{k=1}^{\max\{m,n\}} c_k x^k, & \text{where } c_k = \begin{cases} a_k + b_k & \text{if } k \leq \min\{m,n\} \\ a_k, & \text{if } n \geq k > m \\ b_k, & \text{if } m \geq k > n \end{cases} \\ & \text{if } a(x) = \displaystyle\sum_{i=1}^{n} a_i x^i \text{ and } b(x) = \displaystyle\sum_{j=1}^{m} b_j x^j \end{cases}$$

and their *multiplication* as the polynomial

$$a(x) \cdot b(x) = \begin{cases} 0, & \text{if } a(x) = 0 \text{ or } b(x) = 0 \\ \displaystyle\sum_{k=1}^{n+m} c_k x^k, & \text{where } c_k = \displaystyle\sum_{j=0}^{k} a_{k-j} \cdot b_j, \\ & \text{if } a(x) = \displaystyle\sum_{i=1}^{n} a_i x^i \text{ and } b(x) = \displaystyle\sum_{j=1}^{m} b_j x^j \end{cases}$$

2.19. **Task for you (Rings of polynomials).** *Let $\mathbb{F}$ be a field. The set*

$$\mathbb{F}[x]$$

together with addition and multiplication from Definition 2.18 is a commutative ring with a unit.[a] *It is called the* ring of polynomials in indeterminate x with coefficients in $\mathbb{F}$.

[a] Recall that a *commutative ring with a unit* is "a field where the invertibility test need not hold", see Definition 2.1.

Hints for the proof. The proof is straightforward. ■

2.20. **Remark (Polynomial rings are integral domains).** Task 2.19 shows that the addition and multiplication of polynomials behave as expected: $\mathbb{F}[x]$ with these operations is a commutative ring with a unit, whenever $\mathbb{F}$ is a field. In fact, more can be said: although $\mathbb{F}[x]$ is *never* a field (for example, the non zero polynomial x in $\mathbb{F}[x]$ can never have an inverse), the ring $\mathbb{F}[x]$ satisfies the following condition:

For all $a(x)$, $b(x)$ in $\mathbb{F}[x]$, the equality $a(x) \cdot b(x) = 0$ holds if and only if $a(x) = 0$ or $b(x) = 0$.

Rings having the above property are called *integral domains*, see, e.g., [43]. Clearly: every field is an integral domain. The ring $\mathbb{Z}$ of integers is an easy example of an integral domain, that is not a field. Of course, $\mathbb{F}[x]$ is an example of an integral domain that is not a field.

See Remark 3.7 below for the whole cornucopia of properties of commutative rings with a unit.

3. Quotients rings of polynomial rings

In this chapter we will deal with a "production" of extensions of fields in a certain uniform way using polynomials. As an example, we will show in Task 3.15 below that the field $\mathbb{C}$ of complex numbers is an extension of the field $\mathbb{R}$ of reals, and that this extension is produced by means of polynomials.

3.1. **Task for you (Congruence modulo a non zero polynomial).** *Let $p(x)$ be a polynomial in $\mathbb{F}[x]$ of degree ≥ 1. Define the following binary relation $\underset{p(x)}{\sim}$ on the set $\mathbb{F}[x]$:*

$$a(x) \underset{p(x)}{\sim} b(x) \quad \textit{iff} \quad a(x) - b(x) = k(x) \cdot p(x) \textit{ for some } k(x) \in \mathbb{F}[x]$$

Then the following conditions hold:

(1) $\underset{p(x)}{\sim}$ *is an equivalence relation on $\mathbb{F}[x]$. That is, the following three conditions hold:*
   (a) $\underset{p(x)}{\sim}$ *is reflexive: $a(x) \underset{p(x)}{\sim} a(x)$ holds for every $a(x)$ in $\mathbb{F}[x]$.*
   (b) $\underset{p(x)}{\sim}$ *is symmetric: if $a(x) \underset{p(x)}{\sim} b(x)$ holds, then so does $b(x) \underset{p(x)}{\sim} a(x)$, for every $a(x)$ and $b(x)$ in $\mathbb{F}[x]$.*
   (c) $\underset{p(x)}{\sim}$ *is transitive: if $a(x) \underset{p(x)}{\sim} b(x)$ and $b(x) \underset{p(x)}{\sim} c(x)$ hold, then so does $a(x) \underset{p(x)}{\sim} c(x)$, for every $a(x)$, $b(x)$ and $c(x)$ in $\mathbb{F}[x]$.*

(2) $\underset{p(x)}{\sim}$ *is a congruence w.r.t. addition and multiplication on $\mathbb{F}[x]$. That is, the following two conditions hold:*
   (a) *The relation $\underset{p(x)}{\sim}$ respects addition: suppose $a(x) \underset{p(x)}{\sim} a'(x)$ and $b(x) \underset{p(x)}{\sim} b'(x)$. Then $a(x) + b(x) \underset{p(x)}{\sim} a'(x) + b'(x)$ holds, for every $a(x)$, $a'(x)$, $b(x)$ and $b'(x)$ in $\mathbb{F}[x]$.*
   (b) *The relation $\underset{p(x)}{\sim}$ respects multiplication: suppose $a(x) \underset{p(x)}{\sim} a'(x)$ and $b(x) \underset{p(x)}{\sim} b'(x)$. Then $a(x) \cdot b(x) \underset{p(x)}{\sim} a'(x) \cdot b'(x)$ holds, for every $a(x)$, $a'(x)$, $b(x)$ and $b'(x)$ in $\mathbb{F}[x]$.*

Hints for the proof. The proof is really straightforward, use the definition of $\underset{p(x)}{\sim}$. ■

3.2. **Task for you (The quotient ring modulo a non zero polynomial).** *Consider the equivalence relation $\underset{p(x)}{\sim}$ from Task 3.1. Denote the quotient set $\mathbb{F}[x]/\underset{p(x)}{\sim}$ by*

$$\mathbb{F}[x]/p(x)$$

and define the operations $\underset{p(x)}{+}$ and $\underset{p(x)}{\cdot}$ on equivalence classes as follows:

$$\begin{aligned} [a(x)]_{\underset{p(x)}{\sim}} \underset{p(x)}{+} [b(x)]_{\underset{p(x)}{\sim}} &= [a(x) + b(x)]_{\underset{p(x)}{\sim}} \\ [a(x)]_{\underset{p(x)}{\sim}} \underset{p(x)}{\cdot} [b(x)]_{\underset{p(x)}{\sim}} &= [a(x) \cdot b(x)]_{\underset{p(x)}{\sim}} \end{aligned}$$

Then $\mathbb{F}[x]/p(x)$ is a commutative ring with a unit, called the quotient of $\mathbb{F}[x]$ modulo $p(x)$.

HINTS FOR THE PROOF. Use Task 3.1 to show that the definition of $\underset{p(x)}{+}$ and $\underset{p(x)}{\cdot}$ is correct, i.e., that it does not depend on the choice of representatives. That $\mathbb{F}[x]/p(x)$ is a commutative ring with a unit now follows immediately. ■

3.3. **Remark (Relaxing the notation in quotient rings modulo a polynomial).** We will relax the heavy notation of Task 3.2 as follows:

(1) We will write

$$a(x) = a'(x) \quad \text{in } \mathbb{F}[x]/p(x)$$

instead of

$$[a(x)]_{\underset{p(x)}{\sim}} = [a'x]_{\underset{p(x)}{\sim}} \quad \text{in } \mathbb{F}[x]/p(x)$$

(2) We will write

$$a(x) + b(x) \quad \text{in } \mathbb{F}[x]/p(x) \qquad \text{and} \qquad a(x) \cdot b(x) \quad \text{in } \mathbb{F}[x]/p(x)$$

instead of

$$[a(x)]_{\underset{p(x)}{\sim}} \underset{p(x)}{+} [b(x)]_{\underset{p(x)}{\sim}} \quad \text{in } \mathbb{F}[x]/p(x) \qquad \text{and} \qquad [a(x)]_{\underset{p(x)}{\sim}} \underset{p(x)}{\cdot} [b(x)]_{\underset{p(x)}{\sim}} \quad \text{in } \mathbb{F}[x]/p(x)$$

Thus, for example, we have

$$7x^2 + 6x + 4 = 6x - 3 \quad \text{in } \mathbb{R}[x]/(x^2+1)$$

and

$$(5x+4) \cdot (-5x+4) = 34 \quad \text{in } \mathbb{R}[x]/(x^2+1)$$

Recall that we have defined irreducible *quadratic* polynomials in Definition 2.6. We will now extend this definition to cover polynomials of any degree.

3.4. **Definition (Irreducible polynomials — Take 2).** A non-zero polynomial $p(x)$ in $\mathbb{F}[x]$ is called *irreducible*, if it cannot be written as a product $a(x) \cdot b(x)$, where both $a(x)$ and $b(x)$ have a smaller degree than $p(x)$.

Let us characterise irreducible polynomials of degree up to 2. For polynomials of degree 2 we also reconcile Definitions 2.6 and 3.4 — irreducibility in the sense of Definitions 2.6 is the same thing as irreducibility in the sense of Definition 3.4. We give the characterisation of irreducible polynomials of degree 3 in Example 3.10 below.

3.5. **Example (Irreducible polynomials of degree up to 2).** Let $\mathbb{F}$ be any field.

(1) The polynomial $p(x) = a$ of degree 0 is not irreducible, since $\mathbb{F}$ is a field: we have $a \neq 0$ and any factorisation $a = a(x) \cdot b(x)$ into polynomials of smaller degree would force $a(x) = 0$ or $b(x) = 0$. But then $a(x) \cdot b(x) = a \neq 0$ would hold, which is impossible in a field.
(2) Any polynomial $p(x) = ax + b$ in $\mathbb{F}[x]$ of degree 1 is clearly irreducible.
(3) Suppose $p(x) = ax^2 + bx + c$ in $\mathbb{F}[x]$ has degree 2. We seem to have two different concepts of irreducibility of $p(x)$: Definition 2.6 and Definition 3.4. We show now that this is not the case.

Thus, the following are equivalent:
(a) The polynomial $p(x)$ is irreducible in the sense of Definition 2.6.
(b) The polynomial $p(x)$ is irreducible in the sense of Definition 3.4.
(a) implies (b). Suppose that $p(x)$ is irreducible in the sense of Definition 2.6 but one can find $a(x)$, $b(x)$ such that both $a(x)$, $b(x)$ have degrees ≤ 1 and the equation $p(x) = a(x) \cdot b(x)$ holds. Clearly, it must be the case that both $a(x)$ and $b(x)$ must be of degree precisely 1. But then the equality

$$ax^2 + bx + c = \underbrace{(Ax + A')}_{=a(x)} \cdot \underbrace{(Bx + B')}_{=b(x)}$$

holds in $\mathbb{F}[x]$, where both $A \neq 0$, $B \neq 0$. This means that the equation $ax^2 + bx + c = 0$ has a solution in $\mathbb{F}$; a contradiction.

(b) implies (a). This direction is trivial.

3.6. **Remark (Irreducible polynomials are "primes" in $\mathbb{F}[x]$).** Recall that a natural number p is a *prime*, if it cannot be written as a product $a \cdot b$, where both a and b are natural numbers smaller than p. In that sense, we can see that Definition 3.4 introduces the concept of "primeness" in the ring $\mathbb{F}[x]$. The only difference is that we measure the "largeness" of $a(x)$ and $b(x)$ in the product $a(x) \cdot b(x)$ by degree rather than by value.

In fact, there is a theory of "abstract primeness" that can be developed in any ring $\mathbb{K}$:

> An element p in a ring $\mathbb{K}$ is called *irreducible*, if from $p = a \cdot b$ it follows that a or b are invertible.

The concept of being prime and being an irreducible polynomial are then special cases of irreducible elements in the integral domain $\mathbb{Z}$ and integral domains of the form $\mathbb{F}[x]$, where $\mathbb{F}$ is a field.

Another important feature of primes is the unique factorisation of numbers into product of primes. Such a theory can be developed in integral domains that are called *unique factorisation domains*. One can prove that, every field is a unique factorisation domain, and, given a unique factorisation domain $\mathbb{K}$, that $\mathbb{K}[x]$ is a unique factorisation domain again. See, e.g., [23]. Thus, for example, $\mathbb{F}[x]$ is a unique factorisation domain, for any field $\mathbb{F}$. Hence, every polynomial in $\mathbb{F}[x]$ can be factored essentially uniquely into a product of irreducible polynomials.

3.7. **Remark (More about various types of rings).** Considerations about "primes" from Remark 3.6 lead to important classes of commutative rings with a unit that are commonly studied in Abstract Algebra. Namely, there are notions

$$\{\text{fields}\} \subsetneqq \{\text{Euclidean domains}\} \subsetneqq \{\text{principal ideal domains}\} \subsetneqq$$

$$\subsetneqq \{\text{unique factorisation domains}\} \subsetneqq \{\text{integral domains}\} \subsetneqq \{\text{commutative rings with a unit}\}$$

that reflect what algebraists study, given a "well-behaved" addition and multiplication. We comment only on Euclidean domains and principal ideal domains, sinve we have not encountered them. Unique factorisation domains were mentioned in Remark 3.6.

(1) A *Euclidean domain* is a commutative ring with a unit $\mathbb{K}$ where one can perform the *division with a remainder*. More in detail, $\mathbb{K}$ is an Euclidean domain, if for every $r \neq 0$ in $\mathbb{K}$ there is a natural number $N(r)$ (called the *norm* of r), and such that for any two a, b in $\mathbb{K}$ with $b \neq 0$ there exist unique q and r in $\mathbb{K}$ such that

$$a = q \cdot b + r, \quad \text{with either } r = 0 \text{ or } N(r) < N(b)$$

For example, $\mathbb{Z}$ is a Euclidean domain: put $N(r) = |r|$. Or, $\mathbb{F}[x]$ is a Euclidean domain with $N(p(x))$=the degree of $p(x)$, whenever $\mathbb{F}$ is a field, as we will prove in Task 3.8 below.

(2) A *principal ideal domain* is a commutative ring with a unit, where every ideal is principal. But what does it mean, really?

An *ideal* in a commutative ring with a unit $\mathbb{K}$ is a subset I of $\mathbb{K}$ that has the following three properties:

(a) If a, b are in I, then $a + b$ is in I.
(b) If a, b are in I, then $a \cdot b$ is in I.
(c) If a is in I, then $r \cdot a$ is in I for every r in $\mathbb{K}$.

And an ideal I is *principal*, if there exists a in $\mathbb{K}$ such that $I = \{r \cdot a \mid r \text{ in } \mathbb{K}\}$.

All of the inclusions above are proper, see [9] for examples.

We will show in what follows that the quotient ring $\mathbb{F}[x]/p(x)$ is a field if and only if $p(x)$ is an irreducible polynomial in $\mathbb{F}[x]$. This is similar to the well-known fact that $\mathbb{Z}_p$ is a field if and only if p is a prime number. Similarly to the case of integers, the key ingredient for proving the result is the division with a remainder for polynomials.

3.8. **Task for you (Division with a remainder in $\mathbb{F}[x]$).** *Suppose that $a(x)$, $b(x)$ are polynomials in $\mathbb{F}[x]$, $b(x) \neq 0$. Then there are unique polynomials $q(x)$, $r(x)$ in $\mathbb{F}[x]$ such that the equality*

$$a(x) = q(x) \cdot b(x) + r(x) \quad \textit{where the degree of } r(x) \textit{ is smaller than the degree of } b(x)$$

holds in $\mathbb{F}[x]$.[a]

[a] The polynomial $q(x)$ is called the *quotient* of $a(x)$ divided by $b(x)$ and the polynomial $r(x)$ is called the *remainder* $a(x)$ divided by $b(x)$.

Hints for the proof.

(1) The existence of $q(x)$ and $r(x)$.

Let n be the degree of $a(x)$ and m the degree of $b(x)$. Since $b(x)$ is a non zero polynomial, we have $m \geq 0$. We will proceed by induction on n. We may suppose that $n \geq m$, since, in case when $n < m$, the existence of $q(x)$ and $r(x)$ is trivial: $q(x) = 0$ and $r(x) = a(x)$.

(a) Suppose $n = 0$. Then $m = 0$. Therefore $a(x) = a_0 \neq 0$ and $b(x) = b_0 \neq 0$. Define $r(x) = 0$ and $q(x) = a_0 \cdot (b_0)^{-1}$.

(b) Suppose that the assertion holds for all polynomials $a'(x)$ with degree n' such that $0 \leq n' < n$. Define

$$a'(x) = a(x) - (a_n \cdot b_m^{-1}) \cdot x^{n-m} \cdot b(x)$$

Then $a'(x)$ has degree smaller than n, since the coefficient at x^n is zero. By induction assumption we have $a'(x) = b(x) \cdot q'(x) + r'(x)$ for some $q'(x)$ a $r'(x)$, where the degree of $r'(x)$ is smaller than m. Then we have

$$\begin{aligned} a(x) &= a'(x) + (a_n \cdot b_m^{-1}) \cdot x^{n-m} \cdot b(x) \\ &= b(x) \cdot q'(x) + r'(x) + (a_n \cdot b_m^{-1}) \cdot x^{n-m} \cdot b(x) \\ &= b(x) \cdot (q'(x) + (a_n \cdot b_m^{-1}) \cdot x^{n-m}) + r'(x) \end{aligned}$$

and we can define $q(x) = (q'(x) + (a_n \cdot b_m^{-1}) \cdot x^{n-m})$ and $r(x) = r'(x)$.

(2) The uniqueness of $q(x)$ and $r(x)$.

Suppose

$$a(x) = b(x) \cdot q_1(x) + r_1(x) = b(x) \cdot q_2(x) + r_2(x)$$

Then $b(x) \cdot (q_1(x) - q_2(x)) = r_2(x) - r_1(x)$. Were the polynomials $q_1(x)$, $q_2(x)$ different, we would have the degree of $b(x) \cdot (q_1(x) - q_2(x))$ at most equal to the degree of $b(x)$.

Since the degree of $r_2(x) - r_1(x)$ is smaller than the degree of $b(x)$, this is a contradiction.

Therefore $q_1(x) = q_2(x)$ and $r_1(x) = r_2(x)$.

The proof is finished. ■

3.9. **Remark.** The proof of Task 3.8 gives the usual recursive algorithm for the polynomial division. Consider the following example.

We are given the polynomials $a(x) = 2x^3 - 4x + 1$ and $b(x) = 3x + 2$ in $\mathbb{R}[x]$, and we are to divide $a(x)$ by $b(x)$ with a remainder. We find the quotient $q(x)$ and the remainder $r(x)$ recursively:

(1) We divide the leading coefficients: $\frac{2}{3}$. That is the leading coefficient of the first approximation: $\frac{2}{3} \cdot x^{3-1} = \frac{2}{3}x^2$.

(2) We multiply $(3x+2) \cdot \frac{2}{3}x^2 = 2x^3 + \frac{4}{3}x^2$ and we compute the correction $(2x^3 - 4x + 1) - (2x^3 + \frac{4}{3}x^2) = -\frac{4}{3}x^2 - 4x + 1$.

(3) Now we are to compute $(-\frac{4}{3}x^2 - 4x + 1) : (3x + 2)$. Approximation: $-\frac{4}{9}x$, correction: $-\frac{28}{9}x + 1$.

(4) We are to compute $(-\frac{28}{9}x + 1) : (3x + 2)$. Approximation: $-\frac{28}{27}$, correction: $\frac{83}{27}$. We stop the algorithm, since the degree of $\frac{83}{27}$ is smaller than the degree of $3x + 2$.

Hence

$$q(x) = \frac{2}{3}x^2 - \frac{4}{9}x - \frac{28}{27} \qquad r(x) = \frac{83}{27}$$

and the equality

$$(2x^3 - 4x + 1) = \Big(\frac{2}{3}x^2 - \frac{4}{9}x - \frac{28}{27}\Big) \cdot (3x + 2) + \frac{83}{27}$$

holds in $\mathbb{R}[x]$.

We often write down the previous computations as

$$
\begin{array}{rrrrl}
\big(2x^3 & & -4x & +1\big) & : \big(3x+2\big) = \frac{2}{3}x^2 - \frac{4}{9}x - \frac{28}{27} + \dfrac{\frac{83}{27}}{3x+2} \\
-2x^3 & -\frac{4}{3}x^2 & & & \\
\cline{1-2}
 & -\frac{4}{3}x^2 & -4x & & \\
 & \frac{4}{3}x^2 & +\frac{8}{9}x & & \\
\cline{2-3}
 & & -\frac{28}{9}x & +1 & \\
 & & \frac{28}{9}x & +\frac{56}{27} & \\
\cline{3-4}
 & & & \frac{83}{27} &
\end{array}
$$

Division with a remainder allows us to complete the characterisation of irreducible polynomials up to degree 3. Recall that we have characterised irreducible polynomials up to degree 2 in Example 3.5.

3.10. **Example (Irreducible polynomials of degree 3).** Let $\mathbb{F}$ be any field. Consider a polynomial $p(x)$ in $\mathbb{F}[x]$ of degree 3. Then the following are equivalent:

(a) $p(x)$ is irreducible in $\mathbb{F}[x]$.
(b) The equation $p(x) = 0$ has no solution in $\mathbb{F}$.

(a) implies (b). If there were a in $\mathbb{F}$ such that $p(a) = 0$, perform the division with a remainder and obtain

$$p(x) = (x - a) \cdot q(x) + r(x), \quad \text{the degree of } r(x) \text{ smaller than the degree of } x - a$$

This means that $r(x) = r$ for some r in $\mathbb{F}$. Since $p(a) = 0$, we must have $r = 0$. Thus,

$$p(x) = (x - a) \cdot q(x)$$

which contradicts the irreducibility of $p(x)$.

(b) implies (a). Suppose $p(x)$ is not irreducible and write $p(x) = a(x) \cdot b(x)$, where both $a(x)$ and $b(x)$ have degrees smaller than 3. Clearly, one of $a(x)$, $b(x)$ must have degree 1. Without loss of generality, assume that $a(x) = Ax + B$, with $a \neq 0$. Define $a = -B \cdot A^{-1}$. Then $p(a) = 0$, which is a contradiction with Condition (b).

3.11. **Example (Irreducibile polynomials of degree ≥ 4).** For polynomials of degree ≥ 4 irreducibility becomes a tricky matter. Clearly, for an irreducible polynomial $p(x)$ (of any degree) in $\mathbb{F}[x]$, the equation $p(x) = 0$ cannot have a solution in $\mathbb{F}$, since that would contradict the irreducibility of $p(x)$ — the argument is the same as in "(a) implies (b)" of Example 3.10. On the other hand, there exist polynomials $p(x)$ that are *not* irreducible, yet the equation $p(x) = 0$ has no solutions in $\mathbb{F}$.

For example, the polynomial

$$p(x) = x^4 + 2x^2 + 1 \quad \text{in } \mathbb{R}[x]$$

of degree 4 is clearly not irreducible in $\mathbb{R}[x]$, since the equality

$$p(x) = (x^2 + 1) \cdot (x^2 + 1)$$

holds in $\mathbb{R}[x]$. But, equally clearly, there is no a in $\mathbb{R}$ such that $p(a) = 0$ holds.

Division with a remainder in $\mathbb{Z}$ allowed us to choose *canonical* representatives in quotient rings of the form $\mathbb{Z}_m$. Indeed, one can write

$$\mathbb{Z}_m = \{0, 1, \ldots, m - 1\}$$

That is, $\mathbb{Z}_m$ is the ring of *remainders* after dividing by m. Analogously, we show now that

$$\mathbb{F}[x]/p(x) = \{a(x) \mid \text{the degree of } a(x) \text{ is smaller than the degree of } p(x)\}$$

That is, the ring $\mathbb{F}[x]/p(x)$ is the ring of all *remainders* after dividing by $p(x)$.

3.12. **Corollary.** *Every element in the quotient ring $\mathbb{F}[x]/p(x)$ has a canonical representative by a polynomial of smaller degree than $p(x)$.*

Hints for the proof. Let $a(x)$ be any element of $\mathbb{F}[x]$. The assertion is trivial for $a(x) = 0$. If $a(x)$ is nonzero, then by Task 3.8 we have

$$a(x) = q(x) \cdot p(x) + r(x) \quad \text{in } \mathbb{F}[x]$$

where the degree of $r(x)$ is smaller than the degree of $p(x)$. Thus, we have the equality

$$a(x) = r(x) \quad \text{in } \mathbb{F}[x]/p(x)$$

We have shown that every element has a canonical representative by a polynomial of smaller degree than $p(x)$. ∎

Similarly to the fact that $\mathbb{Z}_p$ is a field if and only if p is a prime, we can characterise rings $\mathbb{F}[x]/p(x)$ which are fields by *irreducibility* of the polynomial $p(x)$.

3.13. **Task for you (When are quotient rings modulo a polynomial fields?).** *Let $p(x)$ be any polynomial in $\mathbb{F}[x]$ of degree ≥ 1. Then the following are equivalent:*

(1) *The ring $\mathbb{F}[x]/p(x)$ is a field.*
(2) *The polynomial $p(x)$ is irreducible in $\mathbb{F}[x]$.*

Hints for the proof. (1) implies (2). Suppose that $p(x)$ is not irreducible and suppose $\mathbb{F}[x]/p(x)$ is a field. Then the equality $p(x) = a(x) \cdot b(x)$ in $\mathbb{F}[x]$ yields the equality

$$0 = a(x) \cdot b(x) \quad \text{in } \mathbb{F}[x]/p(x)$$

which contradicts the fact that $\mathbb{F}[x]/p(x)$ is a field.

(2) implies (1). Let $a(x)$ be a non zero element in $\mathbb{F}[x]/p(x)$. We want to prove that $(a(x))^{-1}$ exists.

Without loss of generality we can assume that the degree n of $a(x)$ is smaller than the degree of $p(x)$, see Corollary 3.12. We proceed by induction on n.

(i) $n = 0$. Then $a(x) = a_0 \neq 0$ and $(a(x))^{-1} = a_0^{-1}$.
(ii) Assume that for all non zero polynomials $a'(x)$ of degree smaller than n the inverse $(a'(x))^{-1}$ exists.

Divide $p(x)$ by $a(x)$ with a remainder:

$$p(x) = q(x) \cdot a(x) + r(x) \quad \text{in } \mathbb{F}[x]$$

Since $p(x)$ is irreducible, it must be the case that $r(x)$ is nonzero.

Proceeding further, we have the equality

$$0 = q(x) \cdot a(x) + r(x) \quad \text{in } \mathbb{F}[x]/p(x)$$

Since the degree of $r(x)$ is smaller than the degree of $a(x)$, the inverse $(-r(x))^{-1}$ of $-r(x)$ exists by induction assumption. Hence we have

$$1 = \underbrace{(-r(x))^{-1} \cdot q(x)}_{=(a(x))^{-1}} \cdot a(x) \quad \text{in } \mathbb{F}[x]/p(x)$$

and the induction step is completed.

∎

3.14. **Task for you (Quotient rings modulo an irreducible polynomial are field extensions).** *Suppose $\mathbb{F}$ is a field and let $p(x)$ be an irreducible polynomial in $\mathbb{F}[x]$ of degree $n \geq 1$. Then*

$$\mathbb{F} \preceq \mathbb{F}[x]/p(x) \quad \textit{and} \quad [\mathbb{F}[x]/p(x) : \mathbb{F}] = n$$

hold.

Hints for the proof. To prove $\mathbb{F} \preceq \mathbb{F}[x]/p(x)$, observe that $\mathbb{F}[x]/p(x)$ is a field by Task 3.13. Thus, it only suffices to show that $\mathbb{F}$ is a subfield of $\mathbb{F}[x]/p(x)$. But this is trivial: all the operations in $\mathbb{F}$ coincide with those in $\mathbb{F}[x]/p(x)$ for polynomials of degree ≤ 0.

To prove $[\mathbb{F}[x]/p(x) : \mathbb{F}] = n$, observe that the set $\{1, x, \ldots, x^{n-1}\}$ forms a basis of $\mathbb{F}[x]/p(x)$ as a vector space over $\mathbb{F}$. ∎

3.15. **Task for you (The complex numbers).** *The field $\mathbb{C}$ of complex numbers is the quotient ring*

$$\mathbb{R}[x]/(x^2+1)$$

Therefore $\mathbb{R} \preceq \mathbb{C}$ and $[\mathbb{C} : \mathbb{R}] = 2$ hold.

Hints for the proof. It is easy to see that any element $a(x)$ in $\mathbb{R}[x]/(x^2+1)$ has a canonical representative of the form $ax+b$, where a, b are reals. This follows from the fact that $a(x)$ can be divided by x^2+1 with a remainder in $\mathbb{R}[x]$. Thus, the equality

$$a(x) = q(x) \cdot (x^2+1) + (ax+b)$$

holds in $\mathbb{R}[x]$ for unique polynomials $q(x)$ and $ax+b$, see Task 3.8. Therefore we obtain the equality

$$a(x) = ax + b \quad \text{in } \mathbb{R}[x]/(x^2+1)$$

by the definition of the relation $\underset{x^2+1}{\sim}$. This fact is an instance of Corollary 3.12.

When working with the above canonical representatives, we can see that

$$\begin{array}{rcll} (ax+b)+(a'x+b) & = & (a+a')x+(b+b') & \text{in } \mathbb{R}[x]/(x^2+1) \\ (ax+b)\cdot(a'x+b) & = & (aa')x^2+(ab+a'b')x+(bb') = (ab+a'b')x+(bb'-aa') & \text{in } \mathbb{R}[x]/(x^2+1) \end{array}$$

where, in case of multiplication, we computed the remainder of when divided by x^2+1: the equality

$$(aa')x^2+(ab+a'b')x+(bb') = (aa')\cdot(x^2+1) + \Big((ab+a'b')x+(bb'-aa')\Big)$$

holds in $\mathbb{R}[x]$.

Since x^2+1 is irreducible in $\mathbb{R}[x]$, the quotient ring $\mathbb{R}[x]/(x^2+1)$ is a field. It is now easy to see that under the identification

$$(ax+b) \leftrightsquigarrow (b+ai) \qquad \text{where } a,\, b \text{ are in } \mathbb{R}$$

the field $\mathbb{R}[x]/(x^2+1)$ is the field $\mathbb{C}$ of complex numbers.

The properties $\mathbb{R} \preceq \mathbb{C}$ and $[\mathbb{C} : \mathbb{R}] = 2$ follow from Task 3.14. ■

As Task 3.15 suggests, we may expect to find roots of the irreducible polynomial $p(x)$ in the field $\mathbb{F}[x]/p(x)$. After all, the number i in $\mathbb{C}$ is the root of x^2+1. This is indeed a general phenomenon as we will show now. To understand this phenomenon, we return to the clumsy but precise notation of Task 3.2 (for the next result only).

3.16. **Task for you (Roots of $p(x)$ in $\mathbb{F}[x]/p(x)$).** *Suppose $p(x)$ is an irreducible polynomial in $\mathbb{F}[x]$. The element $\vartheta = [x]_{\underset{p(x)}{\sim}}$ in $\mathbb{F}[x]/p(x)$ is the root in $\mathbb{F}[x]/p(x)$ of the polynomial $p(x)$ with coefficients in $\mathbb{F}[x]/p(x)$.*

Hints for the proof. Suppose

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0 \quad \text{where } a_n,\, \ldots,\, a_0 \text{ are in } \mathbb{F} \text{ and } a_n \neq 0$$

holds. Since $\mathbb{F} \preceq \mathbb{F}[x]/p(x)$ holds, we can treat $p(x)$ in $\mathbb{F}[x]$ as a polynomial with coefficients in the field $\mathbb{F}[x]/p(x)$. Indeed: $p(x)$, as a polynomial with coefficients in the field $\mathbb{F}[x]/p(x)$, is the expression

$$[a_n]_{\underset{p(x)}{\sim}} \cdot x^n + [a_{n-1}]_{\underset{p(x)}{\sim}} \cdot x^{n-1} + \ldots + [a_0]_{\underset{p(x)}{\sim}} \quad \text{where } a_n,\, \ldots,\, a_0 \text{ are in } \mathbb{F} \text{ and } a_n \neq 0$$

Given any element $[a(x)]_{\underset{p(x)}{\sim}}$ in $\mathbb{F}[x]/p(x)$, we can *evaluate* the above polynomial in $[a(x)]_{\underset{p(x)}{\sim}}$ by computing the sum

$$p\Big([a(x)]_{\underset{p(x)}{\sim}}\Big) = [a_n]_{\underset{p(x)}{\sim}} \underset{p(x)}{\cdot} [a(x)]^n_{\underset{p(x)}{\sim}} \underset{p(x)}{+} [a_{n-1}]_{\underset{p(x)}{\sim}} \underset{p(x)}{\cdot} [a(x)]^{n-1}_{\underset{p(x)}{\sim}} \underset{p(x)}{+} \ldots \underset{p(x)}{+} [a_0]_{\underset{p(x)}{\sim}}$$

in $\mathbb{F}[x]/p(x)$.

Therefore

$$\begin{aligned} p(\vartheta) &= [a_n]_{\underset{p(x)}{\sim}} \underset{p(x)}{\cdot} \vartheta^n \underset{p(x)}{+} [a_{n-1}]_{\underset{p(x)}{\sim}} \underset{p(x)}{\cdot} \vartheta^{n-1} \underset{p(x)}{+} \ldots \underset{p(x)}{+} [a_0]_{\underset{p(x)}{\sim}} \\ &= [a_n]_{\underset{p(x)}{\sim}} \underset{p(x)}{\cdot} [x]^n_{\underset{p(x)}{\sim}} \underset{p(x)}{+} [a_{n-1}]_{\underset{p(x)}{\sim}} \underset{p(x)}{\cdot} [x]^{n-1}_{\underset{p(x)}{\sim}} \underset{p(x)}{+} \ldots \underset{p(x)}{+} [a_0]_{\underset{p(x)}{\sim}} \\ &= [a^n_x + a_{n-1}x^{n-1} + \ldots + a_0]_{\underset{p(x)}{\sim}} \\ &= [p(x)]_{\underset{p(x)}{\sim}} \\ &= [0]_{\underset{p(x)}{\sim}} \end{aligned}$$

holds in $\mathbb{F}[x]/p(x)$. ■

Notice that the above is in perfect accordance with the fact that

$$i = [x]_{\underset{x^2+1}{\sim}}$$

is the root in $\mathbb{C}$ of the polynomial $x^2 + 1$, regarded as a polynomial with *complex* coefficients.

4. The Hermite-Lindemann-Weierstrass Theorem and transdendency of e and π

Some irrational numbers a are roots of polynomial equations $p(x) = 0$, where $p(x)$ is in $\mathbb{Q}[x]$. For example, $a = \sqrt[3]{2}$ is clearly the root of the polynomial $p(x) = x^3 - 2$ in $\mathbb{Q}[x]$. There are irrational numbers that we "suspect" *not* to be roots of any polynomial with rational coefficients. One of our "prime suspects" is the irrational number π, giving the ratio of the circumference of a circle to its diameter. That π is not a root of any polynomial with rational coefficients is indeed the case and it is a highly nontrivial result. We show in this chapter how this property of π can be established. We will see that as a consequence of another highly nontrivial result, called *The Hermite-Lindemann-Weierstrass Theorem* that we formulate (but not prove) in Theorem 4.6 below. The theorem deals with *algebraicity* and *transcendency* of some real numbers over the field of rationals. We start with a general definition of algebraic and transcendental elements of a general field.

4.1. **Definition (Algebraic and transcendental elements).** Suppose $\mathbb{F}_1$, $\mathbb{F}_2$ are fields such that $\mathbb{F}_1 \preceq \mathbb{F}_2$ holds. An element α in $\mathbb{F}_2$ is called *algebraic* over $\mathbb{F}_1$, if there is a polynomial $p(x) \in \mathbb{F}_1[x]$ such that $p(\alpha) = 0$. An element α in $\mathbb{F}_2$ is called *transcendental* over $\mathbb{F}_1$, if it is not algebraic over $\mathbb{F}_1$.

4.2. **Remark (Minimal polynomial).** Suppose $\mathbb{F}_1$, $\mathbb{F}_2$ are fields such that $\mathbb{F}_1 \preceq \mathbb{F}_2$ holds. If α in $\mathbb{F}_2$ is algebraic over $\mathbb{F}_1$, then the set

$$D = \{n \in \mathbb{N} \mid n \text{ is the degree of } p(x) \text{ in } \mathbb{F}_1[x] \text{ such that } p(\alpha) = 0\}$$

is a non-empty subset of $\mathbb{N}$. Hence D has a minimal element, say n_0. Any polynomial $p(x)$ in $\mathbb{F}_1[x]$ of degree n_0 such that $p(\alpha) = 0$ is called a *minimal polynomial* of α. Observe that every minimal polynomial is irreducible.

4.3. **Task for you.** *Suppose $\mathbb{F}_1$, $\mathbb{F}_2$ are fields such that $\mathbb{F}_1 \preceq \mathbb{F}_2$ holds. Suppose that α in $\mathbb{F}_2$ is algebraic over $\mathbb{F}_1$. Then $\mathbb{F}_1(\alpha)$ over $\mathbb{F}_1$ has $\{1, \alpha, \ldots, \alpha^{n-1}\}$ as a basis, where n is the degree of a minimal polynomial for α. In particular, $[\mathbb{F}_1(\alpha) : \mathbb{F}_1] = n$ holds.*

Hints for the proof. Clearly, $\mathbb{F}_1(\alpha)$ is essentially the same as the field $\mathbb{F}_1[x]/p(x)$, where $p(x)$ is a minimal polynomial for α. ■

4.4. **Task for you (Algebraic elements form a field).** *Suppose* $\mathbb{F}_1$, $\mathbb{F}_2$ *are fields such that* $\mathbb{F}_1 \preceq \mathbb{F}_2$ *holds. Then the set*

$$\mathbb{A} = \{\alpha \in \mathbb{F}_2 \mid \alpha \textit{ is algebraic over } \mathbb{F}_1\}$$

forms a field such that

$$\mathbb{F}_1 \preceq \mathbb{A} \preceq \mathbb{F}_2$$

holds.

Hints for the proof. Consider elements α, β in $\mathbb{F}_2$, that are algebraic over $\mathbb{F}_1$. Consider further the extensions $\mathbb{F}_1 \preceq \mathbb{F}_1(\alpha) \preceq \mathbb{F}_1(\alpha, \beta)$. Since the equality

$$[\mathbb{F}_1(\alpha, \beta) : \mathbb{F}_1] = [\mathbb{F}_1(\alpha, \beta) : \mathbb{F}_1(\alpha)] \cdot [\mathbb{F}_1(\alpha) : \mathbb{F}_1]$$

holds by Task 2.12, the extension $\mathbb{F}_1 \preceq \mathbb{F}_1(\alpha, \beta)$ is finite. Thus, all the elements of $\mathbb{F}_1(\alpha, \beta)$ are algebraic over $\mathbb{F}_1$. Hence $\alpha + \beta$, $\alpha - \beta$, $\alpha\beta$ and α/β (if $\beta \neq 0$) are algebraic over $\mathbb{F}_1$.

Thus, $\mathbb{A}$ is a field, such that $\mathbb{F}_1 \preceq \mathbb{A} \preceq \mathbb{F}_2$ holds. ■

4.5. **Task for you (The existence of reals transcendental over rationals).** *Denote by*

$$\mathbb{A} = \{\alpha \in \mathbb{R} \mid \alpha \textit{ is algebraic over } \mathbb{Q}\}$$

the field of reals that are algebraic over rationals. Then the set

$$\mathbb{R} \setminus \mathbb{A}$$

is non empty. Hence there exist reals that are transcendental over rationals.

Hints for the proof. The proof we give will be highly non-constructive.[2] In fact, we will show that the set $\mathbb{A}$ is countable. Since $\mathbb{R}$ is an uncountable set, then the set $\mathbb{R} \setminus \mathbb{A}$ must be uncountable as well. In particular, the set $\mathbb{R} \setminus \mathbb{A}$ of all reals transcendental over rationals must be non empty.

Observe that

$$\mathbb{A} = \bigcup_{n \in \mathbb{N}} K_n$$

where K_n consists of roots of all polynomials in $\mathbb{Q}^{\leq n}[x]$. Since every polynomial in $\mathbb{Q}^{\leq n}[x]$ has at most n roots in $\mathbb{R}$ and since every $\mathbb{Q}^{\leq n}[x]$ is a countable set, the set K_n is countable. Thus, $\mathbb{A}$ is a countable union of countable sets, hence $\mathbb{A}$ is countable. ■

The proof of Task 4.5 does not exhibit even one transcendental number, although we know from the proof that "almost every garden-variety real number must be transcendental". The first explicit transcendental real was constructed by Joseph Liouville[3] in [30]. He showed that the real number, given by the sum of the (obviously convergent) infinite series

$$\sum_{k=1}^{+\infty} \frac{1}{10^{k!}} = 0.110001000000000000000001\ldots$$

is transcendental.

There are various other reals that we "suspect" to be transcendental over rationals. For example, the *Euler's number* e and the *Ludolph's number* π "do not seem" to be roots of polynomials with rational coefficients. That this is indeed the case follows from a highly non-trivial result proved — in a series of papers [20], [28] and [45] — by Charles Hermite, Ferdinand von Lindemann and Karl Weierstrass.[4]

[2] The proof we give is exactly the proof given by Georg Cantor in [6]. Observe that the proof uses (the countable version of) the Axiom of Choice.

[3] Joseph Liouville (24 March 1809 – 8 September 1882) was a French mathematician. He worked in many areas of mathematics, including number theory, differential geometry and topology.

[4] Charles Hermite (24 December 1822 – 14 January 1901) was a French mathematician, working mainly in number theory and theory of elliptic functions. Ferdinand von Lindemann (12 April 1852 – 6 March 1939) was a German mathematician. He is best known for proving transdendency of π in 1882. Lindeman was a PhD advisor of David Hilbert and Hermann Minkowski. Karl Theodor Wilhelm Weierstrass (31 October 1815 – 19 February 1897) was a German mathematician. He is often called the father of modern analysis.

4.6. **Theorem (Hermite-Lindemann-Weierstrass Theorem).** *Let $\mathbb{A}$ be the field of all numbers, algebraic over rationals. Suppose that $\alpha_1, \ldots, \alpha_n$ are pairwise distinct elements of $\mathbb{A}$. Then the set*

$$\{e^{\alpha_1}, \ldots, e^{\alpha_n}\}$$

is linearly independent over $\mathbb{A}$.

The proof of the above theorem uses various slightly advanced methods from calculus of the real line and therefore we do not give it. We refer to, e.g., [33] for the full proof.

4.7. **Corollary (Transcendency of e and π).** *The real numbers e and π are transcendental over rationals.*

Hints for the proof. Use Hermite-Lindemann-Weierstrass Theorem 4.6 to conclude that e is transcendental. Then use the formula $e^{i\pi} = -1$ to conclude that $i\pi$ cannot be algebraic, hence π cannot be algebraic. ■

As we will see in Chapters 6 and 8 below, transcendency of π is the obstacle that prevents us from *squaring the circle* by means of straightedge and compass or by means of origami (i.e., by means of paper folding).

5. The field of constructible numbers

In this chapter we will introduce an ordered field $\mathbb{E}$ of *constructible numbers* such that

$$\mathbb{Q} \preceq \mathbb{E} \preceq \mathbb{R} \qquad \text{and} \qquad \mathbb{E} \neq \mathbb{R}$$

holds, and we will characterise elements of $\mathbb{E}$ as precisely those real numbers r, such that the *segment of length* $|r|$ is constructible by straightedge and compass.

Axioms for Straightedge & Compass Constructions

(SC1) Given points A, B, one can draw a line ℓ through A, B, extending it indefinitely in each direction.

(SC2) Given points A, B, one can draw a line segment AB connecting A and B.

(SC3) Given a point P and a line segment AB, one can draw a circle with the centre at P of radius being the length of AB.

The intersections of constructed lines and circles can also be constructed:

(C1) The point of intersection of two distinct lines can be constructed.

(C2) The points of intersection (provided that they exist) of a circle and a line can be constructed.

(C3) The points of intersection (provided that they exist) of two circles can be constructed.

We have already constructed the midpoint of a segment in Example 1.1 by the above rules. We now arm ourselves with the basic constructions that we will use later.

5.1. **Task for you (The basic constructions with compass and straightedge).** *Using compass and straightedge, we can:*

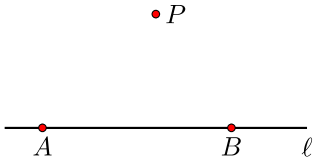
(1) *Given a line ℓ and a point P not on ℓ, construct a line perpendicular to ℓ and passing through P.*

(2) *Given a line ℓ and a point P not on ℓ, construct a line parallel to ℓ and passing through P.*

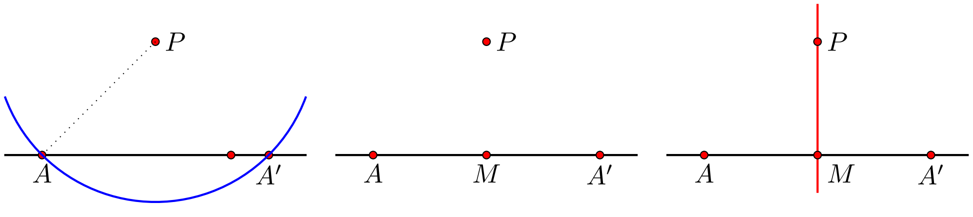
(3) *Given two lines non-parallel lines ℓ_1 and ℓ_2, construct the bisector of the angle between ℓ_1 and ℓ_2.*

Hints for the proof.

(1) Choose any two different points A, B on the line ℓ.



Proceed by the following steps.



Step 1 Step 2 Step 3

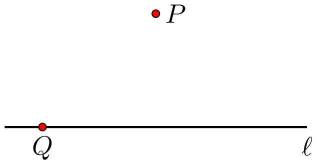
(Step 1): Draw a circle with the centre at P with radius PA. Denote by A' the intersection of this circle with line AB.[5]

(Step 2): Construct the midpoint M of segment AA' as it was done in Example 1.1.

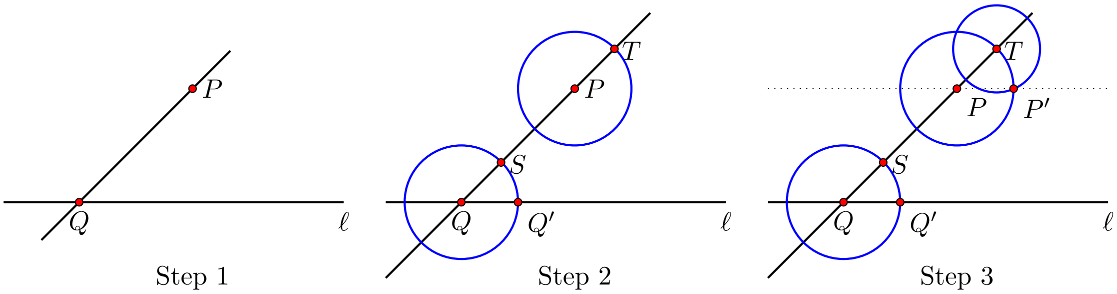
(Step 3): Draw the line through M and P.

Then the line MP is perpendicular to ℓ.

(2) Choose any point Q on the line ℓ, "to the left of P".[6]



Proceed by the following steps.



Step 1 Step 2 Step 3

(Step 1): Construct the line through P, Q.

(Step 2): Draw any circle with the centre at Q. Then draw the circle of *the same radius* with centre in P. Denote the intersections ("in the direction of P") of these circles with line PQ by S and T, respectively. Denote by Q' the intersection of ℓ and the first circle ("in the direction of P").

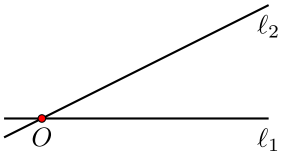
(Step 3): Construct the circle with the centre at T of radius SQ'. Denote the point of intersection "in the direction of Q'" by P'.

The line through P, P' is parallel to the line ℓ.

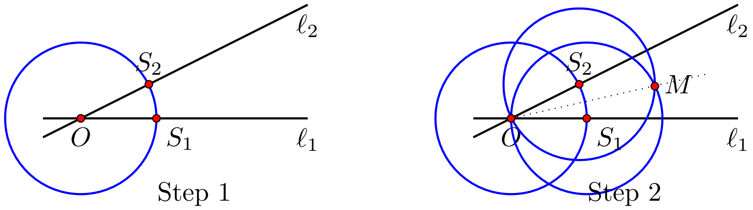
[5] Such a point A' does not exist, if the line AB is a tangent of the circle with centre at P and radius PA. Since the points A and B are different, notice that at least one of the circles with center P and radii PA, PB, respectively, will yield a point A' that we need.

[6] The construction works as well with choosing a point "to the right of P".

(3) Consider two non-parallel lines ℓ_1 and ℓ_2, denote by O their intersection.



Proceed by the following steps.



Step 1

Step 2

(Step 1): Draw any circle with the centre at 0. Denote the intersections of this circle with line ℓ_1, ℓ_2 by S_1, S_2, respectively.

(Step 2): Draw the circles with the centres at S_1 and S_2, respectively, with the same radii as before. Denote by M the point of intersection of these circles "opposite O".

The line through O and M is the bisector of the angle.

The proof is finished. ■

5.2. **Definition (Numbers constructible by straightedge and compass).** We say that a real number r is *constructible by straightedge and compass*, if the segment of length $|r|$ is constructible using axioms (SC1)–(SC3) and rules (C1)–(C3).

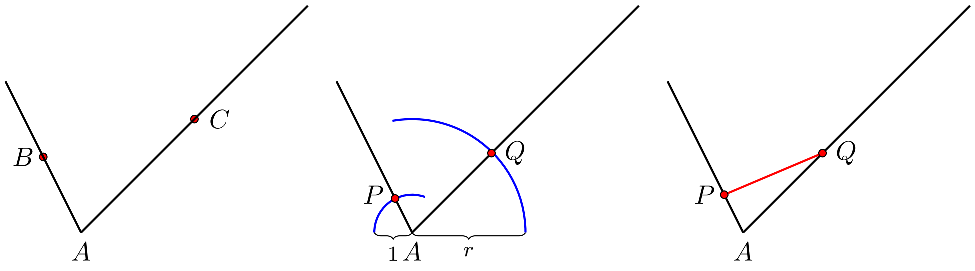
5.3. **Task for you.** *Suppose real numbers* 1, r, s *are constructible by straightedge and compass. Then the real numbers*

$$r+s \qquad r-s, \ (\textit{if } s<r) \qquad rs \qquad r^{-1}, \ (\textit{if } r\neq 0)$$

are constructible by straightedge and compass.

Hints for the proof. The assertion is trivial for $r+s$ and $r-s$, if $s<r$.

For rs, consider the following steps



Step 1

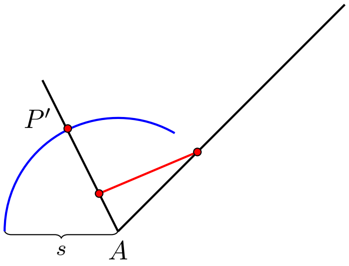
Step 2

Step 3

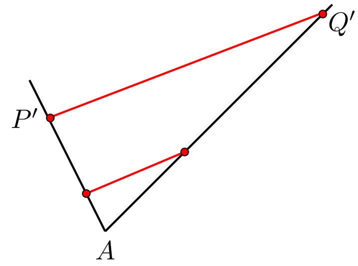
(Step 1): Choose any three points A, B, C not on one line and draw half-lines AB and AC using the straightedge.

(Step 2): Use compass to draw circles with centre at A and of radii 1 and r, respectively. Denote the intersections with AB and AC by P and Q, respectively.

(Step 3): Draw the segment PQ using straightedge.



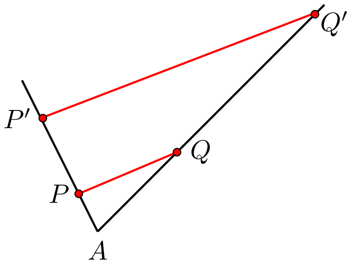
Step 4



Step 5

- **(Step 4):** Using compass, draw a circle with centre at A and radius s. Denote the intersection of AB and this circle by P'.
- **(Step 5):** Using compass and straightedge, draw a line parallel to PQ, passing through P'. Use Task 5.1 to accomplish that. Denote the intersection of AC and this line by Q'.

We have ended up with the picture

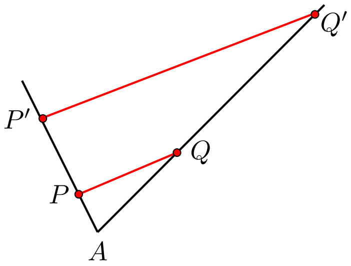


where the triangles $\triangle APQ$ and $\triangle AP'Q'$ are congruent. Hence the we have equalities

$$\frac{s}{1} = \frac{|AP'|}{|AP|} = \frac{|AQ'|}{|AQ|} = \frac{|AQ'|}{r}$$

proving that the length $|AQ'|$ is rs.

To prove that r^{-1} is constructible, use arguments similar to the above, and construct congruent triangles



with $|AP'| = 1$, $|AQ| = 1$, and $|AQ'| = r$. Then

$$\frac{1}{r} = \frac{|AQ|}{|AQ'|} = \frac{|AP|}{|AP'|} = \frac{|AP|}{1}$$

proves that $|AP| = r^{-1}$. ■

5.4. **Corollary (Field of constructible numbers).** *The constructible real numbers form a subfield of $\mathbb{R}$. We denote this subfield by $\mathbb{E}$. Moreover*

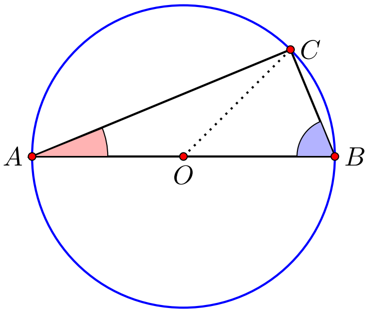
$$\mathbb{Q} \preceq \mathbb{E} \preceq \mathbb{R}$$

holds.

5.5. **Remark (Two more classical results).** In Task 5.6 we will construct a *square root* of a constructible length. The construction hinges upon two classical results of Greek geometry, that we recall now.

The first result is due to *Thales of Miletus*.[7]

(1) *In a triangle inscribed into a circle with one side being the circle's diameter, the angle opposite to the diameter is a right angle.*[8]



The triangle $\triangle AOC$ is isosceles and so is the triangle $\triangle OBC$. Thus, the equalities

$$\measuredangle CAO = \measuredangle OCA \quad \text{and} \quad \measuredangle OBC = \measuredangle OCB$$

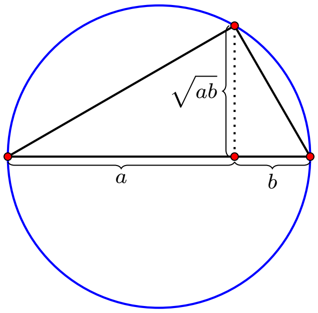
hold. Therefore the equalities

$$\begin{aligned} 2\cdot\measuredangle ACB &= 2\cdot(\measuredangle OCA + \measuredangle OCB) \\ &= 2\cdot\measuredangle OCA + 2\cdot\measuredangle OCB \\ &= (180^\circ - \measuredangle AOC) + (180^\circ - \measuredangle COB) \\ &= 360^\circ - (\measuredangle AOC + \measuredangle COB) \\ &= 180^\circ \end{aligned}$$

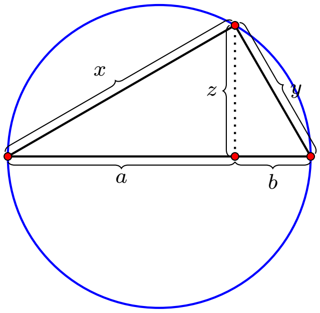
hold and they show that $\measuredangle ACB = 90^\circ$, as desired.

The second result[9] comes as Corollary to Proposition VI.8 in [11].

(2) *In a triangle inscribed into a circle with one side being the circle's diameter, the altitude opposite to the diameter is the geometric mean of the pieces of the hypotenuse.*



Let us denote the remaining two segments by x and y and the length of the altitude by z:



Then the equalities

$$x^2 + y^2 = (a+b)^2 \qquad a^2 + z^2 = x^2 \qquad b^2 + z^2 = y^2$$

[7] Thales of Miletus (cca 640BC–cca 546BC) was a Greek mathematician, astronomer and philosopher. Aristotle regarded him as the first philosopher in the Greek tradition.

[8] See [11], Proposition III.31.

[9] The result will be used not only in Task 5.6, but also in giving *squaring of a rectangle* in Task 6.7.

hold by Pythagoras' Theorem, since all three triangles are right-angled. By substituting the last two equalities into the first one, we obtain

$$a^2 + b^2 + 2z^2 = (a+b)^2$$

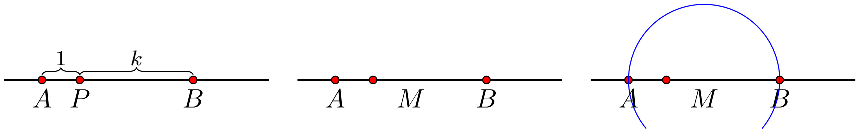
from which the desired equality

$$z^2 = ab$$

follows immediately.

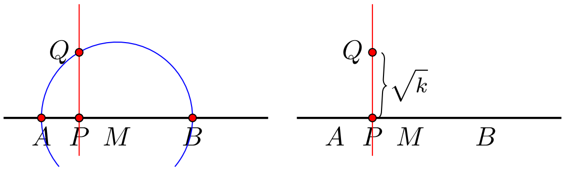
5.6. **Task for you.** *Suppose that k is in $\mathbb{E}$. Then $\sqrt{k}$ is in $\mathbb{E}$.*

Hints for the proof. Consider the following construction:



Step 1 Step 2 Step 3

(Step 1): Draw any line and, using compass, mark three points A, P, B on it, such that $|AP| = 1$ and $|PB| = k$.
(Step 2): Find the midpoint M of the segment AB as in Example 1.1.
(Step 3): Draw a circle with centre at M and radius $|AM|$.



Step 4 Step 5

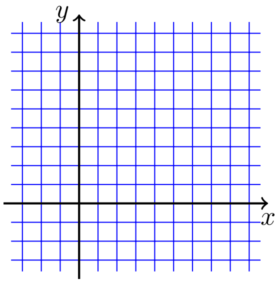
(Step 4): Draw a line, perpendicular to the line AB, passing through point P. Denote by Q the intersection of this line and the circle.
(Step 5): The length of the segment PQ is $\sqrt{k}$ by Remark 5.5.

■

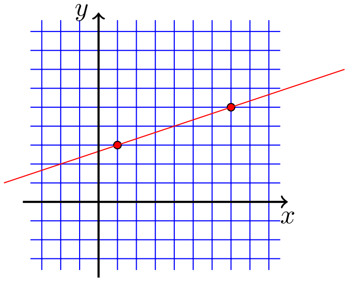
5.7. **Definition (Planes, lines, and circles of a subfield of $\mathbb{R}$).** Suppose $\mathbb{F} \preceq \mathbb{R}$. Then

(1) A *plane of* $\mathbb{F}$ consists of all points (x, y) in $\mathbb{R} \times \mathbb{R}$ such that both $x \in \mathbb{F}$ and $y \in \mathbb{F}$.
(2) A *line in the plane of* $\mathbb{F}$ is any line passing through two points of the plane of $\mathbb{F}$.
(3) A *circle in the plane of* $\mathbb{F}$ is any circle with centre in the plane of $\mathbb{F}$, such that at least one point of the circumference is in the plane of $\mathbb{F}$.

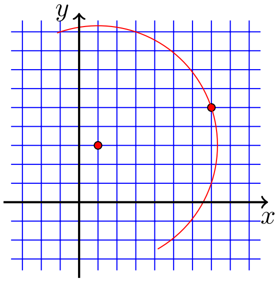
5.8. **Remark.** The plane of $\mathbb{F}$, with $\mathbb{F} \preceq \mathbb{R}$, is simply a lattice of points in the "usual plane" $\mathbb{R} \times \mathbb{R}$, spanned by points that have both coordinates in $\mathbb{F}$. See the following picture:



A line in the plane of $\mathbb{F}$ is any line in $\mathbb{R} \times \mathbb{R}$ that passes through (at least) two points of the lattice above:



A circle in the plane of $\mathbb{F}$ is any circle in $\mathbb{R} \times \mathbb{R}$ that has a centre in the point of a lattice and at least on point of the lattice lies on its circumference:



5.9. **Task for you.** *Let $\mathbb{F} \preceq \mathbb{R}$. Then the following hold:*

(1) *Every line in the plane of $\mathbb{F}$ can be represented by an equation*

$$ax + by + c = 0, \quad \text{where } a,\ b,\ c \text{ are in } \mathbb{F}$$

(2) *Every circle in the plane of $\mathbb{F}$ can be represented by an equation*

$$x^2 + y^2 + ax + by + c = 0 \quad \text{where } a,\ b,\ c \text{ are in } \mathbb{F}$$

Hints for the proof.

(1) From linear algebra we know that an equation of a line passing through points (x_1, y_1), (x_2, y_2) is given by the equality

$$\begin{vmatrix} x & x_1 & x_2 \\ y & y_1 & y_2 \\ 1 & 1 & 1 \end{vmatrix} = 0$$

Expanding the determinant, we obtain the equation

$$\underbrace{(y_1 - y_2)}_{=a\in\mathbb{F}} x + \underbrace{(x_2 - x_1)}_{=b\in\mathbb{F}} y + \underbrace{(x_1y_2 - x_2y_1)}_{=c\in\mathbb{F}} = 0$$

that has the desired form.

(2) Let us have a circle with centre in (x_1, x_2) and let (x_2, y_2) be a point on its circumference. Then the circle has the equation

$$(x - x_1)^2 + (y - y_1)^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$$

which can be simplified to

$$x^2 + y^2 + \underbrace{(-2x_1)}_{=a\in\mathbb{F}} x + \underbrace{(-2y_1)}_{=b\in\mathbb{F}} y + \underbrace{(2x_1x_2 + 2y_1y_2 - x_2^2 - y_2^2)}_{=c\in\mathbb{F}} = 0$$

■

5.10. **Task for you.** *Let $\mathbb{F} \preceq \mathbb{R}$. Then the following hold:*

1. *A point of intersection of two lines in the plane of $\mathbb{F}$ is a point in the plane of $\mathbb{F}$.*
2. *A point of intersection of a line in the plane of $\mathbb{F}$ and a circle in the plane of $\mathbb{F}$ is either a point in the plane of $\mathbb{F}$ or a point in the plane of $\mathbb{F}(\sqrt{k})$ for some k in $\mathbb{F}$.*
3. *A point of intersection of two different circles in the plane of $\mathbb{F}$ is either a point in the plane of $\mathbb{F}$ or a point in the plane of $\mathbb{F}(\sqrt{k})$ for some k in $\mathbb{F}$.*

Hints for the proof.

(1) Consider the system of equations

$$\begin{array}{rcl} a_1x + b_1y + c_1 & = & 0 \\ a_2x + b_2y + c_2 & = & 0 \end{array}$$

If each equation represents a line in the plane of $\mathbb{F}$, then all coefficients above are in $\mathbb{F}$. If the two lines intersect, then the point of intersection is given by Cramer's Rule as

$$x = \frac{\begin{vmatrix} -c_1 & b_1 \\ -c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} \qquad y = \frac{\begin{vmatrix} a_1 & -c_1 \\ a_2 & -c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}$$

which is a point in the plane of $\mathbb{F}$.

(2) The point of intersection of a line and a circle is given by the solution of the system

$$\begin{array}{rcl} x^2 + y^2 + a_1x + b_1y + c_1 & = & 0 \\ a_2x + b_2y + c_2 & = & 0 \end{array}$$

The second equation can be solved for either x or y, since it cannot be the case that both $a_2 = 0$ and $b_2 = 0$. Without loss of generality, let us suppose that we can solve for y:

$$y = (b_2)^{-1} \cdot (-a_2x - c_2)$$

and we can substitute this y to the equation $x^2 + y^2 + a_1x + b_1y + c_1 = 0$. After simplifying, it is clear that we obtain a quadratic equation of the form

$$Ax^2 + Bx + C = 0$$

with coefficients in $\mathbb{F}$. If $A = 0$, then the solution of the above is given by $x = -B^{-1} \cdot C$, which is an element of $\mathbb{F}$. If $A \neq 0$, the solution of the above quadratic equation is given by the formula

$$x_{1,2} = (2A)^{-1} \cdot (-B \pm \sqrt{B^2 - 4AC})$$

where $B^2 - 4AC \geq 0$, since we know a solution must exist. Clearly, both x_1 and x_2 are either in $\mathbb{F}$ or in $\mathbb{F}(\sqrt{k})$ for $k = B^2 - 4AC$.

Substituting x_1 and x_2 to the equation $y = (b_2)^{-1} \cdot (-a_2x - c_2)$, we conclude that the corresponding y_1 and y_2 are either in $\mathbb{F}$ or in $\mathbb{F}(\sqrt{k})$.

We conclude that the points of intersection of a circle in the plane of $\mathbb{F}$ and the line in the plane of $\mathbb{F}$ are either in the plane of $\mathbb{F}$ or in in the plane of $\mathbb{F}(\sqrt{k})$ for some k in $\mathbb{F}$.

(3) The point of intersection of two different circles is given by the solution of the system

$$\begin{array}{rcl} x^2 + y^2 + a_1x + b_1y + c_1 & = & 0 \\ x^2 + y^2 + a_2x + b_2y + c_2 & = & 0 \end{array}$$

Observe that if $a_1 = a_2$ and $b_1 = b_2$, then $c_1 = c_2$ must hold, for otherwise the system would have no solution at all. But if $a_1 = a_2$, $b_1 = b_2$ and $c_1 = c_2$ holds, then both equations describe the same circle. Hence, we have to conclude that $a_1 \neq a_2$ or $b_1 \neq b_2$ must hold.

By subtracting (say) the first equation from the second, we obtain the system

$$\begin{array}{rcl} x^2 + y^2 + a_1x + b_1y + c_1 & = & 0 \\ (a_2 - a_1)x + (b_2 - b_1)y + (c_2 - c_1) & = & 0 \end{array}$$

where $a_2 - a_1 \neq 0$ or $b_2 - b_1 \neq 0$ holds. Now solve the system by the method described in part (2) of this proof.

We conclude that the points of intersection of two different circles in the plane of $\mathbb{F}$ are either in the plane of $\mathbb{F}$ or in in the plane of $\mathbb{F}(\sqrt{k})$ for some k in $\mathbb{F}$.

■

5.11. **Task for you (Characterisation of $\mathbb{E}$).** *For a real number r, the following conditions are equivalent:*

(1) $r \in \mathbb{E}$.

(2) *There exists a sequence $\mathbb{F}_0 \preceq \mathbb{F}_1 \preceq \ldots \preceq \mathbb{F}_N$ of fields such that $\mathbb{F}_0 = \mathbb{Q}$, $r \in \mathbb{F}_N$, and $[\mathbb{F}_{i+1} : \mathbb{F}_i] = 2$, for all $i = 0, \ldots, N-1$.*

Hints for the proof. (1) implies (2). A constructible number is given by repeating axioms (SC1)–(SC3) finitely many times. Since Task 5.6 shows that we are dealing with field extensions of degree up to 2, the assertion follows.

(2) implies (1). We proceed by induction on $N \geq 0$.

(a) If $N = 0$, then r is a rational number, which is constructible.

(b) Suppose $N > 0$ and suppose that any element of $\mathbb{F}_{N-1}$ is constructible. Since $\mathbb{F}_N = \mathbb{F}_{N-1}(\sqrt{k})$ for some k in $\mathbb{F}_{N-1}$, the elements of $\mathbb{F}_N$ are constructible by Task 5.6.

■

5.12. **Corollary.** *If r is a constructible number, then the equality $[\mathbb{Q}(r) : \mathbb{Q}] = 2^A$ holds for some natural number A.*

Hints for the proof. By Task 5.11 there exists a sequence $\mathbb{F}_0 \preceq \mathbb{F}_1 \preceq \ldots \preceq \mathbb{F}_N$ of fields such that $\mathbb{F}_0 = \mathbb{Q}$, $r \in \mathbb{F}_N$, and $[\mathbb{F}_{i+1} : \mathbb{F}_i] = 2$, for all $i = 0, \ldots, N-1$. Hence

$$[\mathbb{F}_N : \mathbb{Q}] = 2^N$$

by Task 2.12. By the same task, the degree $[\mathbb{Q}(r) : \mathbb{Q}]$ divides $[\mathbb{F}_N : \mathbb{Q}]$, since $\mathbb{Q} \preceq \mathbb{Q}(r) \preceq \mathbb{F}_N$ holds. Thus, there exist a natural number A such that $[\mathbb{Q}(r) : \mathbb{Q}] = 2^A$ holds. ■

6. Some classical problems from Ancient Greece

In this chapter we formulate three classical problems known from Greek geometry: given a straightedge and a compass, can one solve the following?

(1) Given a cube with edge of length a, can one construct a cube that has the volume $2a^3$, i.e., *twice* as big as the original cube? This problem is known as *doubling the cube*.

(2) Given a general angle α, can one construct an angle $\alpha/3$? This problem is known as *trisecting an angle*.

(3) Given a disc of radius r, can one construct $r\sqrt{\pi}$, i.e., the edge of a square of the same area as the original disc? This problem is known as *squaring the circle*.

We will show how to answer the above questions with the machinery we developed in the previous chapters. In fact, as we will see, all three questions have to be answered in negative.

Doubling the cube. There is a story behind doubling the cube, which — true or not — is amusing.[10] The god Apollo had sent a plague on the people of Delos Island. The citizens decided to consult the oracle at Delphi for a solution to problems that intensified relationships among the citizens. The oracle responded: "You have to double the shrine of Apollo." It so happened that the shrine had the shape of a cube. Oracle's

[10] See Plato's *Republic*, [36], Book VII.530.

answer seemed strange to Delians and they consulted Plato to clear the oracle's advice. Plato interpreted oracle's answer as the advice of Apollo: "Study geometry and calm down the disputes".

As it turns out, the oracle's advice could not have been carried out, at least not with straightedge and compass. This was proved (much later) by Pierre Laurent Wantzel[11] and we show how it can be proved by the techniques developed in these notes.

6.1. **Task for you (The impossibility of doubling the cube).** *It is impossible to double the cube using compass and straightedge.*

Hints for the proof. Suppose that we are given a cube with edge of length a. We are to construct b such that the equality

$$b^3 = 2a^3$$

holds, or, equivalently, such that the equality

$$b = a\sqrt[3]{2}$$

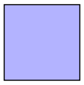
holds.

Obviously, if we construct $\sqrt[3]{2}$ using straightedge and compass, then we double the cube. But, by Corollary 5.12, the number $\sqrt[3]{2}$ is *not* constructible, since

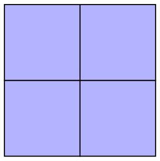
$$[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$$

holds. Indeed, it is easy to see that $\{1, \sqrt[3]{2}, \sqrt[3]{2^2}\}$ is a basis of $\mathbb{Q}(\sqrt[3]{2})$ over $\mathbb{Q}$. See also Task 4.3. ■

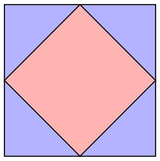
6.2. **Example (Doubling the square).** On the other hand, it is quite easy to *double the square*.[12] To prove it, consider a square as in the picture



and, using compass and straightedge, draw another three copies of it as in the picture



The square having diagonals of those squares as the edges has the area double the area of the original square:



Indeed, if a is the length of the edge of the given square, then we have constructed the length $b = a\sqrt{2}$. Thus, the equality

$$b^2 = \left(a\sqrt{2}\right)^2 = 2a^2$$

holds.

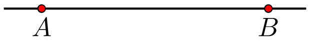
[11] Pierre Laurent Wantzel (5 June 1814 – 21 May 1848) was a French mathematician. In his paper [44] he not only proved that doubling the cube and trisecting an angle cannot be solved by straightedge and compass, he also gave the characterisation of which regular polygons are constructible. He thus showed that the sufficient conditions formulated by Karl Friedrich Gauss are also necessary.

[12] See, for example, the dialogue of *Socrates* with a slave of Meno, paragraphs 82a and further in [35].

Trisecting an angle. Since it is possible to *bisect* an angle using straightedge and compass, the natural question arises: can one *trisect a general angle* using straightedge and compass?

Perhaps surprisingly, the answer is *no*. This may come as a surprise, since we can *trisect a general segment* using straightedge and compass as we show now.[13]

6.3. **Example (Trisecting a segment).** Consider a general segment AB as in the picture

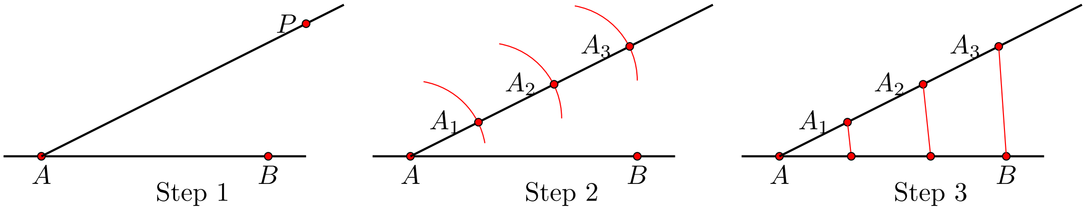


and perform the following steps:

(Step 1): Construct any line passing through A and denote it by AP.

(Step 2): Construct consecutively points A_1, A_2, A_3 as the intersection of circles with a fixed radius and centres at A, A_1 and A_2 consecutively.

(Step 3): Draw the segment A_3B and then draw parallel lines to it through points A_2 and A_1. The intersections of these lines with the segment AB yield the trisection of AB.



Before we show that the trisection of an angle is impossible, we will prove a property of a certain *cubic* equation.

6.4. **Task for you.** *Suppose $\mathbb{F} \preceq \mathbb{R}$. Suppose that the equation $x^3 - 3x - 1 = 0$ has a root in some quadratic extension $\mathbb{F}(\sqrt{k})$ of $\mathbb{F}$. Then $x^3 - 3x - 1 = 0$ has a root in $\mathbb{F}$.*

Hints for the proof. Suppose $r = a + b\sqrt{k}$, where a, b are in $\mathbb{F}$, is a root of $x^3 - 3x - 1 = 0$. Then only one of the two following cases can happen.

(1) $b = 0$. Then $r = a$ is in $\mathbb{F}$ and the proof is finished.

(2) $b \neq 0$. We will prove that in this case, $-2a$ is a root of $x^3 - 3x - 1 = 0$.

Substituting $r = a + b\sqrt{k}$ to $x^3 - 3x - 1 = 0$ we obtain

$$0 = \left(a + b\sqrt{k}\right)^3 - 3\left(a + b\sqrt{k}\right) - 1 = \left(a^3 + 3ab^2k - 3a - 1\right) + \left(3a^2b + b^3k - 3b\right) \cdot \sqrt{k}$$

Since $\{1, \sqrt{k}\}$ is linearly independent over $\mathbb{F}$, it must be the case that

$$\left(a^3 + 3ab^2k - 3a - 1\right) = 0 \quad \text{and} \quad \left(3a^2b + b^3k - 3b\right) = 0$$

Since $b \neq 0$, we can divide the second equation by b. Hence we are dealing with the system

$$\left(a^3 + 3ab^2k - 3a - 1\right) = 0 \quad \text{and} \quad \left(3a^2 + b^2k - 3\right) = 0$$

or, equivalently, with the system

$$\left(a^3 + 3ab^2k - 3a - 1\right) = 0 \quad \text{and} \quad b^2k = 3(1 - a^2)$$

By substituting $b^2k = 3(1 - a^2)$ to the first equation we obtain

$$(-2a)^3 - 3(-2a) - 1 = 0$$

Thus, $-2a$ is an element of $\mathbb{F}$ that solves the equation $x^3 - 3x - 1 = 0$.

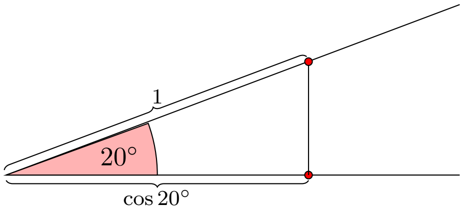
The proof is finished. ■

[13] By an easy generalisation of the procedure in Example 6.3 one can show that one can divide any segment into n equal parts for any natural number $n \geq 2$.

6.5. **Task for you (The impossibility of trisection of an angle).** *It is impossible to trisect a general angle using compass and straightedge.*

Hints for the proof. We will show that the angle 60° cannot be trisected using straightedge and compass. That is, we will show that the angle 20° *cannot* be constructed using compass and straightedge.

The angle 60° can be constructed using compass and straightedge. If that angle could be trisected by straightedge and compass, the angle 20° would be constructible as well. Then one could construct cos 20° by using the right-angled triangle



and therefore the number $2\cos 20°$ could be constructed. We will show that the constructibility of $2\cos 20°$ leads to contradiction.

First notice that, for any α, the equality

$$\cos 3\alpha = 4\cos^3\alpha - 3\cos\alpha$$

holds. Since $\cos 60° = 1/2$, this means that

$$1/2 = 4\cos^3 20° - 3\cos 20°$$

holds, or, equivalently, that the equality

$$8\cos^3 20° - 6\cos 20° - 1 = 0$$

holds. In other words, $2\cos 20°$ must be the root of the cubic equation

$$x^3 - 3x - 1 = 0$$

Using Corollary 5.12 and Task 6.4, it follows that $2\cos 20°$ must be a rational number. But the only possible rational roots[14] of $x^3 - 3x - 1$ are 1 or -1. Since the equalities

$$1^3 - 3\cdot 1 - 1 = -3 \quad\text{and}\quad (-1)^3 - 3\cdot(-1) - 1 = -4$$

hold, the proof is finished. ■

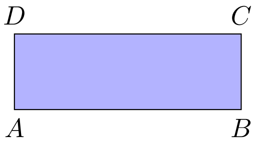
6.6. **Remark (Some angles can be trisected).** It is worth noticing that *some* angles can be trisected using straightedge and compass. For example, one can trisect the right angle, since one can construct the angle of 30° using straightedge and compass.

Squaring of the circle. The problem of squaring the circle requires, given a disk, to construct a square of the same area as the given disc using only straightedge and compass.

We do not have any sources on the history of the above problem. There seems to be no other reason than that a circle (and a disk it circumscribes) is the logical next step of geometric shapes that one may try to square. After all, as we show now, both the rectangles and triangles *can* be squared.

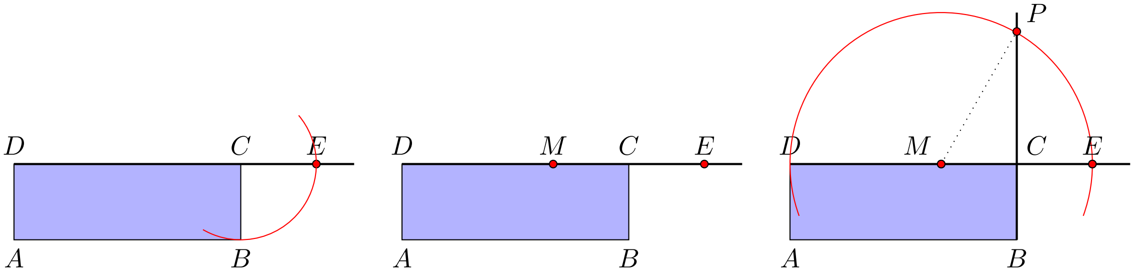
6.7. **Example (Squaring of the rectangle).** Any rectangle can be squared. That is: given a rectangle $ABCD$, one can find a square of the same area as the rectangle.

Suppose that a rectangle



[14] Recall that if the equation $a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0 = 0$ with integer coefficients $a_n, \ldots, a_0$, $a_n \neq 0$, has a rational solution a/b with a, b relatively prime, then a divides a_0 and b divides a_n. This follows from the fact that the equalities $-a_0\cdot b^n = a\cdot(a_n a^{n-1} + a_{n-1}a^{n-2}b + \ldots + a_1 b^{n-1})$ and $-a_n\cdot a^n = b\cdot(a_{n-1}a^{n-2} + a_{n-2}a^{n-3}b + \ldots + a_0 b^{n-1})$ must hold, and from the fact that a, b are relatively prime.

is given. Proceed by the following steps.



Step 1 Step 2 Step 3

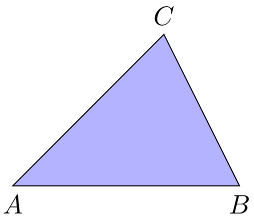
(Step 1): Prolong the segment CD and draw on it point E such that $|CE| = |CB|$ holds.
(Step 2): Find the midpoint of M of the segment DE.
(Step 3): Draw a circle with centre at M and radius ME. Prolong segment CB and denote its intersection with the circle by P. The segment CP is the edge of the desired square.

Indeed, the area of the rectangle $ABCD$ is $|AB| \cdot |BC|$. And we have the following equalities:

$$\begin{aligned} |AB| \cdot |BC| &= |DC| \cdot |BC| = \Big(|DM| + |MC|\Big) \cdot |CE| = \Big(|DM| + |MC|\Big) \cdot \Big(|DM| - |MC|\Big) \\ &= |DM|^2 - |MC|^2 = |PC|^2 \end{aligned}$$

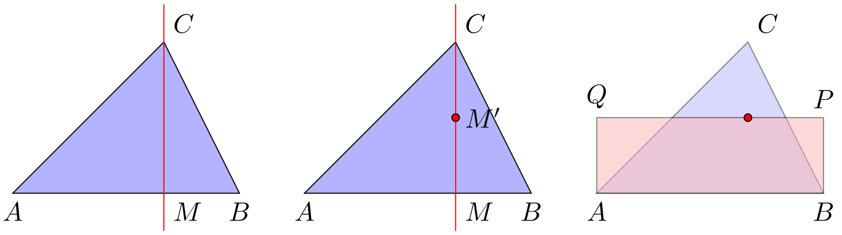
6.8. **Example** (**Squaring of the triangle).** Any triangle can be squared. That is: given a triangle, one can construct a square of the same area as the triangle.

Suppose that a triangle



is given. The idea of squaring the triangle is as follows: first we find a rectangle of the same area as the triangle and then we square the rectangle using Example 6.7.

(Step 1): Draw a line perpendicular to AB, passing through the point C. Denote its intersection with AB by M.
(Step 2): Cut the segment MC in half and denote the resulting midpoint by M'.
(Step 3): Draw a rectangle $ABPQ$, with the line PQ passing through M'.



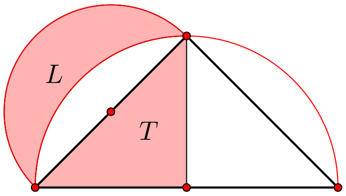
Step 1 Step 2 Step 3

The area of $\triangle ABC$ is $(1/2) \cdot |MC| \cdot |AB|$, which is the same as the area of the rectangle $ABPQ$, since $|BP| = |M'C| = (1/2) \cdot |MC|$. Now find a square of the same area as $ABPQ$ as in Example 6.7.

Squaring rectangles and triangles might sound trivial: both rectangles and triangles are quite "edgy" figures. That one can, in fact, square certain areas not bounded by a polygon therefore came as a surprising and nontrivial result. Let us see the historically first such result.

6.9. **Remark (The lunes of Hippocrates).** The first breakthrough in squaring areas that are not bounded by a polygon was made by *Hippocrates of Chios*.[15] He dealt with squaring areas that are called lunes. A *lune* is a concave region bounded by two circular arcs. In the following picture, consider an isosceles right-angled triangle. Draw a circle with centre at the midpoint of the hypotenuse passing through the third vertex. Then draw a circle with centre on one of the catheti passing through its endpoints. This yields a lune on one of the catheti. Hippocrates proved the following result:

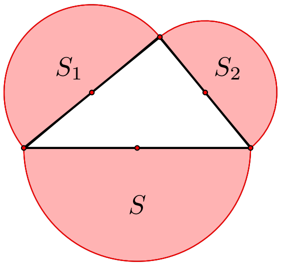
> *In the picture below, the equality $L = T$ holds. That is, the area of the lune is equal to one half of the area of the triangle.*



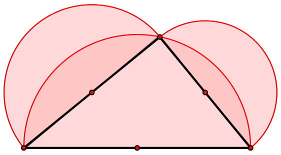
In fact, the above is an easy consequence of a generalisaton of *The Pythagoras' Theorem*:[16]

> *Given a right-angled triangle, then the sum of areas of semicircles at their catheti equals the sum of the area of the semicircle at the hypotenuse.*

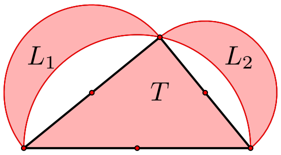
In other words, in the following picture



the equality $S_1 + S_2 = S$ holds. But by drawing the above picture differently, we obtain



We conclude that the sum of the areas L_1, L_2 of the two lunes at the catheti is equal to the area T of the whole triangle, i.e., that the equality $L_1 + L_2 = T$ holds:



Since one can square T using straightedge and compass by Examples 6.8 and 6.7, one can also square the sum $L_1 + L_2$. Hippocrates' result is a special case of this, when the right-angled triangle is isosceles.

All of the above nurtured hopes that we would be able to find a proof that a circle can be squared using compass and straightedge. Hippocrates, of course, did not stop at the case of lunes described above. As of

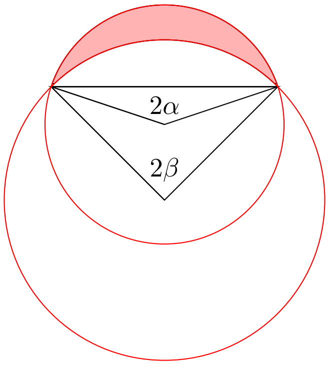
[15] Not to be confused with his contemporary *Hippocrates of Kos* (cca 460BC – cca 370BC), the founder of Greek medicine and the author of physicians' oath. Hippocrates of Chios (470BC–410BC) was the first to write a systematically organised texbook on geometry. His book, called *Elements*, was written some 100 years before Euclid's. However, his book has not survived, the only work by Hippocrates that survived is squaring the lunes. The work was preserved by *Simplicius of Cilicia* (cca 490AD – cca 560AD).

[16] Euclid, of course, proves the "classical" Pythagoras' Theorem in Book I, Proposition I.47. He also proves the "most general form" of Pythagoras' Theorem: *If one erects similar figures with corresponding sides on the sides of a right-angled triangle, then the sum of the areas of the ones on the two smaller sides equals the area of the one on the larger side.* See [11], Book VI, Proposition VI.31. Clearly, the "semicircle" Pythagoras' Theorem is an instance of the "most general form".

roughly 440BC, three types of lunes were known to him to be squarable. Then the progress stopped for over two millenia.

In 1766, *Martin Johann Wallenius*[17] found two more squarable types of lunes, making the total of types of lunes that can be squared to be five.

What are these five types of squarable lunes? Any concave lune can be given by the common chord and by two angles α and β on the same side of the common chord, see the picture



If we restrict ourselves to the case that $\alpha = m\vartheta$ and $\beta = n\vartheta$, where $m > n$ are relatively prime natural numbers, then the following can be proved (see, e.g., [38][18]):

> *A lune with $\alpha = m\vartheta$ and $\beta = n\vartheta$ is squarable if and only if*
>
> $$(m, n) \in \{(2, 1), (3, 1), (3, 2), (5, 1), (5, 3)\}$$
>
> *holds.*

Thus, the five types of squarable lunes are:[19]

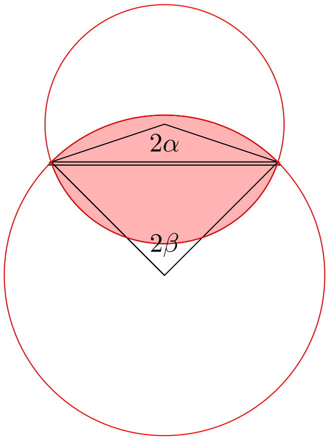
Values (m, n)	$\cos 2\vartheta$	2β	2α	Known as squarable to:
$(2, 1)$	0	$90°$	$180°$	Hippocrates
$(3, 1)$	$\left(\sqrt{3} - 1\right)/2$	$\approx 68.5°$	$\approx 205.6°$	Hippocrates
$(3, 2)$	$\left(\sqrt{33} - 1\right)/8$	$\approx 107.2°$	$\approx 160.9°$	Hippocrates
$(5, 1)$	$\left(\sqrt{5 + 4\sqrt{5}} - 1\right)/4$	$\approx 46.9°$	$\approx 234.4°$	Wallenius
$(5, 3)$	$\left(\sqrt{5} - \sqrt{3} + \sqrt{20 + 2\sqrt{15}}\right)/\left(4\sqrt{3}\right)$	$\approx 100.8°$	$\approx 168°$	Wallenius

[17] Martin Johann Walenius (18 March 1731 – 22 October 1773) was a Swedish mathematician.

[18] The full proof that there are no other squarable lunes than the five types above is in Chapter 16 of [38], where the result is attributed to Soviet mathematicians Nikolai Grigorievich Chebotarev (in 1933: the proof for m, n odd) and Anatoly V. Dorodnov (in 1945: the proof for the rest possible m, n) without any reference to a paper or a book. Let us remark that the proof is rather nontrivial.

[19] The table comes from [38]. See also [37].

How about *convex* lunes? These can be described by a common chord and two angles on opposite sides of the chord as in the following picture:



It can be proved that *no* convex lune can be squarable. See [37] or [38] for the full proof. Notice that the convex lune with $2\alpha = 2\beta = 180°$ is a disk. Hence, in view of Task 6.10 below, nonsquarability of convex lunes is perhaps not that surprising.

6.10. **Task for you (The impossibility of squaring the circle).** *It is impossible to square the circle using compass and straightedge.*

Hints for the proof. A circle of radius r circumscribes the disk of area πr^2. To square a circle, one would have to construct the length $a = r\sqrt{\pi}$. We show that $\sqrt{\pi}$ is not in $\mathbb{E}$, proving thus the impossibility of squaring the circle.

Indeed, were $\sqrt{\pi}$ constructible, then π would be constructible as well. But π is transcendental over $\mathbb{Q}$ by Corollary 4.7. Hence π cannot be constructed by Corollary 5.12. ■

7. Huzita-Hatori axioms for origami

In the late 1980's, an interest rose in constructions that are performed by *folding* a sheet of paper. Such constructions were not new: it is an old Japanese tradition of *origami* art. What was new, however, was the realisation that paper folding — just like straightedge and compass — *constructs lengths*. And that, perhaps, given a set of axioms for paper folding, one would be able to characterise the lengths constructible by paper folding. This is precisely the topic of the current chapter.

We introduce a set of axioms for paper folding that gives rise to the field $\mathbb{O}$ of *origami numbers*. The field $\mathbb{O}$ satisfies

$$\mathbb{Q} \preceq \mathbb{O} \preceq \mathbb{R} \qquad \text{and} \qquad \mathbb{O} \neq \mathbb{R}$$

We also show that

$$\mathbb{E} \preceq \mathbb{O} \qquad \text{and} \qquad \mathbb{E} \neq \mathbb{O}$$

hold. Thus, the origami constructions can be shown to be *stronger* than constructions by straightedge and compass.

Huzita-Hatori Axioms for Origami Constructions

(HH1) Given two distinct points P_1 and P_2, there is a unique fold f that passes through both of them.

(HH2) Given two distinct points P_1 and P_2, there is a unique fold f that places P_1 onto P_2.

(HH3) Given two lines ℓ_1 and ℓ_2, there is a fold f that places ℓ_1 onto ℓ_2.

(HH4) Given a point P and a line ℓ, there is a unique fold f perpendicular to ℓ that passes through point P.

(HH5) Given two points P_1 and P_2 and a line ℓ, there is a fold f that places P_1 onto ℓ and that passes through P_2.

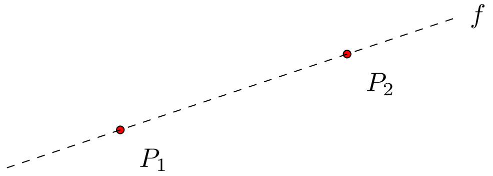
(HH6) Given two points P_1 and P_2 and two lines ℓ_1 and ℓ_2, there is a fold f that places P_1 onto ℓ_1 and P_2 onto ℓ_2.

(HH7) Given one point P and two lines ℓ_1 and ℓ_2, there is a fold f that places P onto ℓ_1 and is perpendicular to ℓ_2.

(C0) We can construct a point as an intersection of two constructible folds using the axioms (HH1)–(HH7) finitely many times.

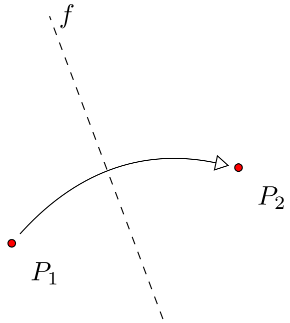
7.1. **Remark** (**Geometric interpretation of the origami axioms).** We illustrate the axioms as follows (folds are shown as dashed lines):

(HH1): Given two distinct points P_1 and P_2, there is a unique fold f that passes through both of them.



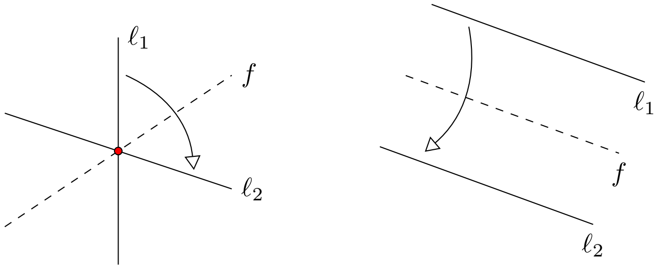
The fold f is the unique line passing through P_1, P_2.

(HH2): Given two distinct points P_1 and P_2, there is a unique fold f that places P_1 onto P_2.



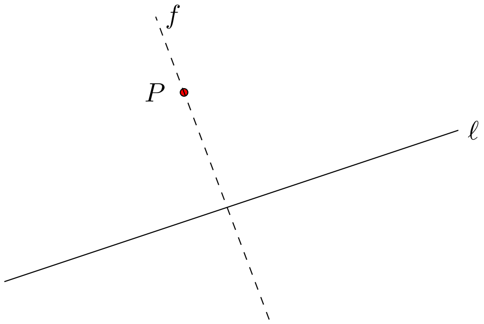
The fold is the perpendicular bisector of the segment P_1P_2.

(HH3): Given two lines ℓ_1 and ℓ_2, there is a fold f that places ℓ_1 onto ℓ_2.



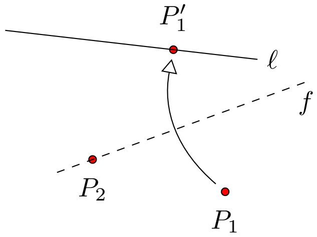
The fold f is the bisector of the angle between ℓ_1 and ℓ_2 in case ℓ_1 and ℓ_2 intersect. If ℓ_1, ℓ_2 are parallel, the fold f is the bisector of ℓ_1 and ℓ_2.

(HH4): Given a point P and a line ℓ, there is a unique fold f perpendicular to ℓ that passes through point P.



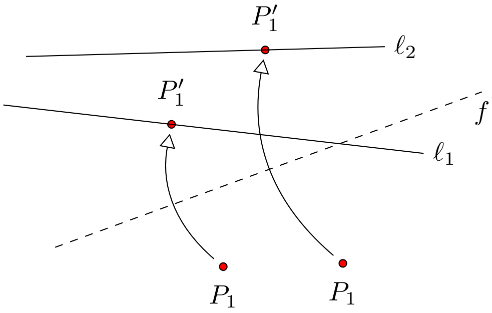
The fold f is the unique line perpendicular to ℓ and passing through P.

(HH5): Given two points P_1 and P_2 and a line ℓ, there is a fold f that places P_1 onto ℓ and that passes through P_2.



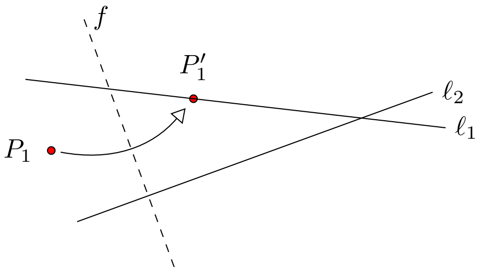
The fold f is the tangent at P_2 to the parabola with focus P_1 and directrix ℓ. We will prove it in Task 7.5 below, after we explain what the directrix and focus of a parabola means in Remark 7.3.

(HH6): Given two points P_1 and P_2 and two lines ℓ_1 and ℓ_2, there is a fold f that places P_1 onto ℓ_1 and P_2 onto ℓ_2.



The fold f is the common tangent to the parabola with focus P_1 and directrix ℓ_1 and to the parabola with focus P_2 and directrix ℓ_2. It is this axiom that will allow us to solve *cubic* equations, see Remark 7.12 and Task 7.13 below.

(HH7): Given one point P and two lines ℓ_1 and ℓ_2, there is a fold f that places P onto ℓ_1 and is perpendicular to ℓ_2.



The fold f is the tangent to the parabola with focus in P_1 and directrix ℓ_1, perpendicular to the line ℓ_2.

7.2. **Remark (The history of axioms (HH1)–(HH7)).** Axioms (HH1)–(HH7) were formulated in the late 1980's by Jacques Justin [24] but the paper went unnoticed. All seven axioms were later independently rediscovered by Humiaki Huzita (the first six) and Koshiro Hatori (axiom (HH7)). Commonly, the set of axioms is referred to as *Huzita-Hatori Axioms*. History is not fair.

Axioms (HH1)–(HH7) are not independent; it has been shown that (HH1)–(HH5) can be done using only (HH6), see [3].

7.3. **Remark (Parabolas, given by a focus and a directrix).** A *parabola* belongs to the class of planar curves that are called *conics*. All conics can be given as the intersection of a plane and a double cone. Therefore sometimes conics are also called *conic sections*. We will not need the full theory of conics, we restrict ourselves to parabolas. We only give some hints to the general theory in Remark 7.4 below. For the modern treatment of conics, we refer, e.g., to the books [25] and [14]. As it turns out, the best milieu to study conics is *complex projective geometry*, and we do not want to go into that here.

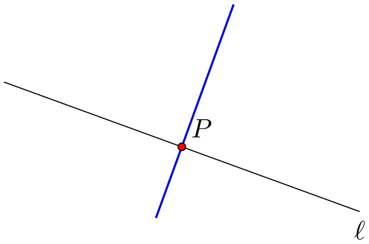
What *is* a parabola, then? The easiest description[20] is as follows: suppose that we are given a line ℓ in the plane and a point P. Consider the set of all points X in the plane such that the distance of X from P is equal to the distance of X from ℓ:

$$\{X \mid \text{dist}(X,P) = \text{dist}(X,\ell)\}$$

Then one of the following two cases happens:

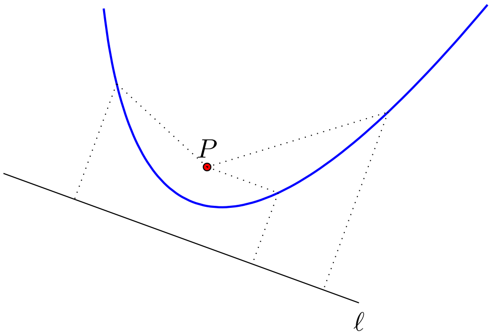
(a) The point P is a point of ℓ.

In this case, the set $\{X \mid \text{dist}(X,P) = \text{dist}(X,\ell)\}$ coincides with the set of points on the line that is perpendicular to ℓ at P. This line is called a *degenerated parabola*.



(b) The point P is not on ℓ.

In this case, all points X lie on the curve that is called a *parabola* with *focus* P and *directrix*[21] ℓ.



That the above definition is equivalent to our usual perception of a parabola as a certain *quadratic curve* is maintained by a little linear algebra, as we show now.

(1) Suppose first that the directrix ℓ has the equation $y + p = 0$ and the point P has coordinates $(0, p)$, where $p > 0$. Then a point X with coordinates (x, y) satisfies $\text{dist}(X,P) = \text{dist}(X,\ell)$ iff

$$y = \frac{1}{4p}x^2$$

Indeed, we have

$$\text{dist}(X,P) = \sqrt{x^2 + (p-y)^2}$$

and

$$\text{dist}(X,\ell) = |y + p|$$

[20] Incidentally, this is the very definition given by Apollonius of Perga.

[21] The word *directrix* means *leader* (as in *leader line*). Since *line* is a femininum in Latin, the adjective acquires the suffix *-ix*, instead of the suffix *-or* for masculina. Thus we have *director* and *directrix* for leaders. Compare with *dominator* and *dominatrix*, *tractor* and *tractrix*, etc.

Thus, the equality

$$\sqrt{x^2+(p-y)^2} = |y+p|$$

which, after taking squares, yields

$$x^2+(p-y)^2 = (y+p)^2$$

and that simplifies to

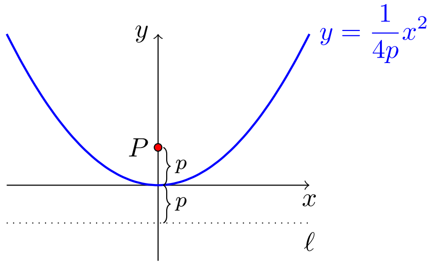
$$x^2-2py = 2py$$

or, equivalently, to

$$y = \frac{1}{4p}x^2$$

as desired.

Thus, in this case, the curve looks like as follows:



(2) In case when ℓ has the equation $ax+by+c=0$ and P is the point with coordinates (p_1,p_2), then a point X with coordinates (x,y) has the distance

$$\operatorname{dist}(X,P) = \sqrt{(p_1-x)^2+(p_2-x)^2}$$

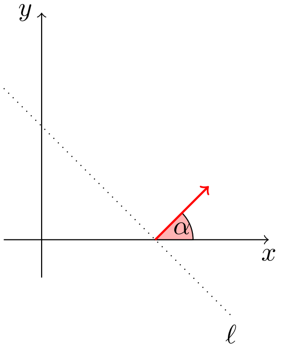
from the point P and the distance of X from ℓ is given by the formula

$$\operatorname{dist}(X,\ell) = \frac{|ax+by+c|}{\sqrt{a^2+b^2}}$$

The above fraction makes sense, since, necessarily, $a^2+b^2 \neq 0$ holds. Thus, without loss of generality, we can assume that $a^2+b^2=1$. and we can define a unique angle α in $[0;2\pi)$ such that the equalities

$$a = \cos\alpha \qquad b = \sin\alpha$$

hold. The angle α is the angle between the *unit normal* of the line $ax+by+c=0$ and the axis x:



The equality $\operatorname{dist}(X,P) = \operatorname{dist}(X,\ell)$ is equivalent to the equality $\operatorname{dist}^2(X,P) = \operatorname{dist}^2(X,\ell)$ and therefore we can write

$$(p_1-x)^2+(p_2-y)^2 = (ax+by+c)^2$$

that can be written as follows:

$$\underbrace{(1-a^2)}_{=b^2=A}\cdot x^2+\underbrace{(-2ab)}_{=2B}\cdot xy+\underbrace{(1-b^2)}_{=a^2=C}\cdot y^2+\underbrace{(-2p_1-2ac)}_{=2D}\cdot x+\underbrace{(-2p_2-2bc)}_{=2E}\cdot y+\underbrace{(p_1^2+p_2^2-c^2)}_{=F}=0$$

We are therefore dealing with the matrix equation

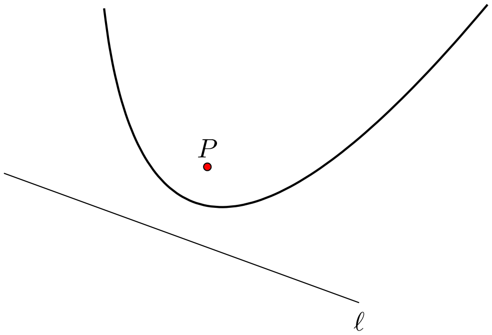
$$\begin{pmatrix} x & y & 1 \end{pmatrix} \cdot \underbrace{\begin{pmatrix} A & B & D \\ B & C & E \\ D & E & F \end{pmatrix}}_{=\mathbf{Q}} \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = 0$$

or, equivalently, with the equation

$$\begin{pmatrix} x & y & 1 \end{pmatrix} \cdot \begin{pmatrix} b^2 & -ab & -p_1 - ac \\ -ab & a^2 & -p_2 - bc \\ -p_1 - ac & -p_2 - bc & p_1^2 + p_2^2 - c^2 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = 0$$

This suggests that we should treat our curve in $\mathbb{R}^3$ as being drawn in the plane with equality $z = 1$.[22]

Let us transform this general case to the case (1) above. The idea is: rotate and translate the general curve



(a) We first rotate by the angle $\beta = \pi/2 - \alpha$. Rotation of the curve in the plane $z = 1$ by angle β is given by the matrix product

$$\begin{pmatrix} \cos\beta & \sin\beta & 0 \\ -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix}^T \cdot \begin{pmatrix} A & B & D \\ B & C & E \\ D & E & F \end{pmatrix} \cdot \begin{pmatrix} \cos\beta & \sin\beta & 0 \\ -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} b & a & 0 \\ -a & b & 0 \\ 0 & 0 & 1 \end{pmatrix}^T \cdot \begin{pmatrix} A & B & D \\ B & C & E \\ D & E & F \end{pmatrix} \cdot \begin{pmatrix} b & a & 0 \\ -a & b & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

since $\cos\beta = \cos(\pi/2 - \alpha) = \sin\alpha = b$ and $\sin\beta = \sin(\pi/2 - \alpha) = \cos\alpha = a$. Therefore, after rotation by β, we obtain the matrix

$$\begin{pmatrix} Ab^2 - 2Bab + Ca^2 & (A-C)ab + B(b^2 - a^2) & Db - Ea \\ (A-C)ab + B(b^2 - a^2) & Aa^2 + 2Bab + Cb^2 & Da + Eb \\ Db - Ea & Da + Eb & F \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & -p_1 b + p_2 a \\ 0 & 0 & -p_1 a - p_2 b - c \\ -p_1 b + p_2 a & -p_1 a - p_2 b - c & p_1^2 + p_2^2 - c^2 \end{pmatrix}$$

(b) After the above step (a), the directrix is parallel to the axis x. We now set the origin of the coordinate system "into the middle" of the focus and the directrix. Translation by a vector

$$\begin{pmatrix} t_1 \\ t_2 \end{pmatrix}$$

[22] Connoisseurs recognise that we are going to treat the curve in the real *affine plane*.

in the plane $z = 1$ is performed by the matrix multiplication

$$\begin{pmatrix}1 & 0 & -t_1\\ 0 & 1 & -t_2\\ 0 & 0 & 1\end{pmatrix}^T \cdot \begin{pmatrix}1 & 0 & Db-Ea\\ 0 & 0 & Da+Eb\\ Db-Ea & Da+Eb & F\end{pmatrix} \cdot \begin{pmatrix}1 & 0 & -t_1\\ 0 & 1 & -t_2\\ 0 & 0 & 1\end{pmatrix} =$$

$$= \begin{pmatrix}1 & 0 & -t_1+Db-Ea\\ 0 & 0 & Da+Eb\\ -t_1+Db-Ea & Da+Eb & t_1^2-2(Db-Ea)t_1-2(Da+Eb)t_2+F\end{pmatrix} =$$

$$= \begin{pmatrix}1 & 0 & -t_1-p_1b+p_2a\\ 0 & 0 & -p_1a-p_2b-c\\ -t_1-p_1b+p_2a & -p_1a-p_2b-c & t_1^2-2(-p_1b+p_2a)t_1-2(-p_1a-p_2b-c)t_2+p_1^2+p_2^2-c^2\end{pmatrix}$$

Put

$$t_1 = -p_1b+p_2a \quad \text{and} \quad t_2 = p_1a+p_2b+c$$

Then our curve has the matrix

$$\begin{pmatrix}1 & 0 & -t_1-p_1b+p_2a\\ 0 & 0 & -p_1a-p_2b-c\\ -t_1-p_1b+p_2a & -p_1a-p_2b-c & t_1^2-2(-p_1b+p_2a)t_1-2(-p_1a-p_2b-c)t_2+p_1^2+p_2^2-c^2\end{pmatrix} =$$

$$= \begin{pmatrix}1 & 0 & 0\\ 0 & 0 & -p_1a-p_2b-c\\ 0 & -p_1a-p_2b-c & -(-p_1b+p_2a)^2+2(p_1a+p_2b+c)^2+p_1^2+p_2^2-c^2\end{pmatrix} =$$

$$= \begin{pmatrix}1 & 0 & 0\\ 0 & 0 & -p_1a-p_2b-c\\ 0 & -p_1a-p_2b-c & (p_1a+p_2b+c)(3p_1a+3p_2b-2c)\end{pmatrix}$$

There are two cases to consider:

(i) $p_1a + p_2b + c = 0$, that is: the focus P lies on the directrix ℓ. Then our curve has the matrix

$$\begin{pmatrix}1 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}$$

and we deal with the *degenerated* parabola $x^2 = 0$.

(ii) $p_1a + p_2b + c \neq 0$, that is: the focus P does not lie on the directrix ℓ. Then we deal with the parabola $(p_1a + p_2b + c)y = x^2 + (p_1a + p_2b + c)(3p_1a + 3p_2b - 2c)$.

7.4. **Remark (General conics).** The classification of general conics by analytical means similar to Remark 7.3 is, of course, possible. It amounts to the classificiation of quadratic equations

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0$$

with real coefficients A, B, C, D, E, F, where $A^2 + C^2 \neq 0$. The classification must be invariant with respect to *rigid motions* in the plane, i.e., w.r.t. to mappings of the form

$$\mathbf{x} \mapsto \mathbf{R}_\alpha \cdot \mathbf{x} + \mathbf{x}_0$$

where $\mathbf{R}_\alpha : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is the rotation counterclockwise by angle α along origin, and where $\mathbf{x}_0$ is a fixed vector in $\mathbb{R}^2$.

The above quadratic equation can be written in matrix form as

$$\mathbf{x}^T \cdot \mathbf{A} \cdot \mathbf{x} + 2\mathbf{b}^T \cdot \mathbf{x} + F = 0$$

where

$$\mathbf{A} = \begin{pmatrix}A & B\\ B & C\end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix}D\\ E\end{pmatrix}$$

or, even more compactly as

$$\begin{pmatrix}\mathbf{x}^T & 1\end{pmatrix} \cdot \begin{pmatrix}\mathbf{A} & \mathbf{b}\\ \mathbf{b}^T & F\end{pmatrix} \cdot \begin{pmatrix}\mathbf{x}\\ 1\end{pmatrix} = 0$$

which expresses the quadratic equation in the plane $z = 1$ in $\mathbb{R}^3$. See the excellent and detailed account of conics and quadrics in, e.g., [14] and [15].

7.5. **Task for you (Geometric interpretation of (HH5), (HH6) and (HH7)).**

(HH5) *Given two points P_1 and P_2 and a line ℓ, there is a fold f that places P_1 onto ℓ and that passes through P_2.*

Then the fold f is the tangent at P_2 to the parabola with focus P_1 and directrix ℓ.

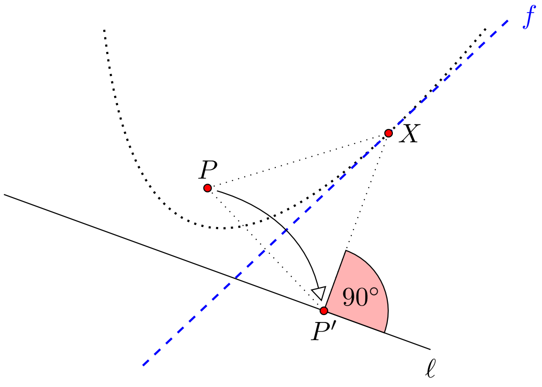
(HH6) *Given two points P_1 and P_2 and two lines ℓ_1 and ℓ_2, there is a fold f that places P_1 onto ℓ_1 and P_2 onto ℓ_2.*

Then the fold f is the common tangent to the parabola with focus P_1 and directrix ℓ_1 and to the parabola with focus P_2 and directrix ℓ_2.

(HH7) *Given one point P and two lines ℓ_1 and ℓ_2, there is a fold f that places P onto ℓ_1 and is perpendicular to ℓ_2.*

Then the fold f is the tangent to the parabola with focus in P_1 and directrix ℓ_1, perpendicular to the line ℓ_2.

Hints for the proof. Axiom (HH5) places P at P' along the fold f, as seen in the picture:



Since the fold f is the perpendicular bisector of the segment PP', any point of f is equidistant to P and P'. If we choose X to have the same distance from ℓ as from P', then the fold f is the tangent at point X to the parabola with focus in P and directrix ℓ.

The interpretations of (HH6) and (HH7) now follow. ■

As in Task 5.1 for compass and straightedge, we can perform the basic constructions using (HH1)–(HH7).

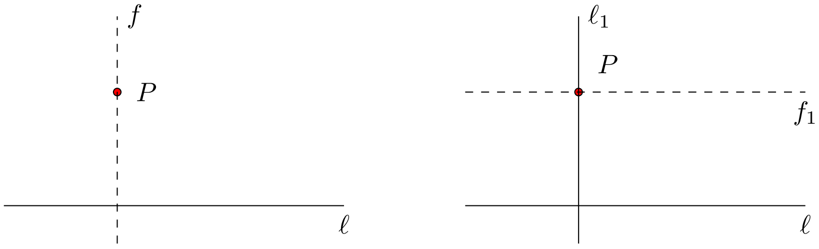
7.6. **Task for you (The basic origami constructions).** *Using axioms* (HH3) *and* (HH4) *for origami, we can:*

(1) *Given a line ℓ and a point P not on ℓ, construct a line perpendicular to ℓ and passing through P.*

(2) *Given a line ℓ and a point P not on ℓ, construct a line parallel to ℓ and passing through P.*

(3) *Given two lines non-parallel lines ℓ_1 and ℓ_2, construct the bisector of the angle between ℓ_1 and ℓ_2.*

Hints for the proof. For (1), use axiom (HH4) and for (3), use axiom (HH3). To prove (2), apply the axiom (HH4) twice: In (Step 1) construct the fold f passing through P and perpendicular to ℓ. Denote the line of the fold by ℓ_1 and in (Step 2) use (HH4) to construct a fold f_2 passing through P and perpendicular to ℓ_1. Denote the line of the fold f_2 by ℓ_2.



Step 1 Step 2

The lines ℓ and ℓ_2 are parallel. ■

7.7. Definition (Origami-constructible numbers). We say that a real number r is *origami-constructible*, if the segment of length $|r|$ is constructible using Axioms (HH1)–(HH7).

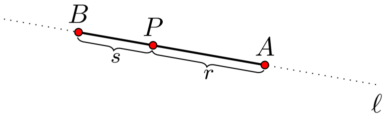
7.8. Task for you. *Suppose real numbers* 1, r, s *are origami-constructible. Then the real numbers*

$$r+s \qquad r-s,\ (\textit{if } s<r) \qquad rs \qquad r^{-1},\ (\textit{if } r\neq 0)$$

are origami-constructible.

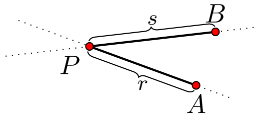
Hints for the proof.

(1) In case we have adjacent segments



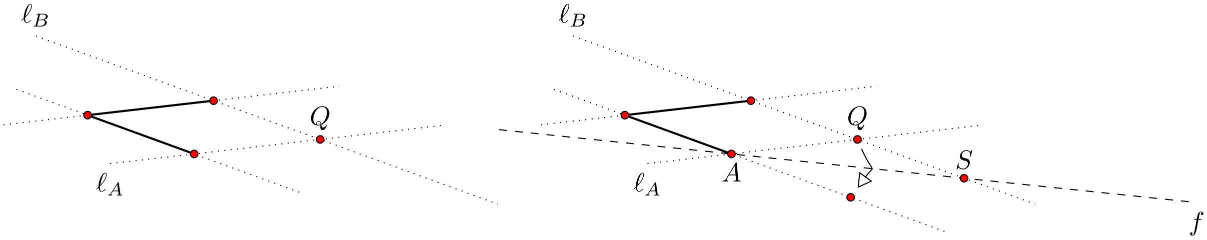
on a common line, it is easy to construct both $r+s$ and $r-s$ (in case $s<r$).

Suppose therefore that we already have the segments



that are not on the same line but they are, without loss of generality, adjacent at one point P.

To construct $r+s$, proceed by the following steps



Step 1 Step 2

Step 1: Using Task 7.6, construct lines ℓ_A and ℓ_B, parallel to PB, PA, and passing through A and B, respectively. Denote by Q the intersection of ℓ_A and ℓ_B.

Step 2: Use (HH3) to construct the fold f that places Q on the line PA and denote the intersection of f with ℓ_B by S. The triangle $\triangle AQS$ is isosceles, hence the length of QS is equal to the length of AQ, which is s.

We now have adjacent segments of lengths r and s, respectively. It is now straighforward that we can construct $r+s$, and $r-s$ (in case $s<r$).

(2) To construct rs and r^{-1}, use the respective constructions from Task 5.3, but using origami axioms to construct parallel lines (as explained in Task 7.6).

■

7.9. **Corollary (Field of origami-constructible numbers).** *The origami-constructible real numbers form a subfield of* $\mathbb{R}$. *We denote this subfield by* $\mathbb{O}$. *Moreover*

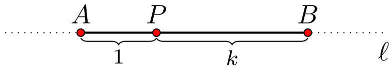
$$\mathbb{Q} \preceq \mathbb{O} \preceq \mathbb{R}$$

holds.

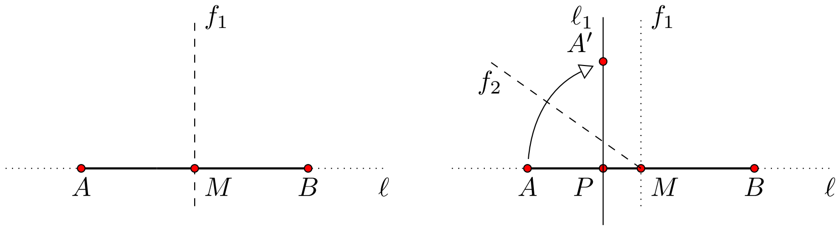
Before we embark on the study of *cubic roots* of origami-constructible numbers, we show that the field $\mathbb{O}$ is closed under square roots. Notice that the method of the proof is the same as that in Task 5.6. We will thus be able to conclude that the field $\mathbb{O}$ of origami-constructible numbers is an extension of the field $\mathbb{E}$ of numbers constructible by straightedge and compass. See Corollary 7.11.

7.10. **Task for you.** *Suppose that* k *is in* $\mathbb{O}$. *Then* $\sqrt{k}$ *is in* $\mathbb{O}$.

Hints for the proof. Suppose that lengths 1 and k have already been constructed, as in the picture



Proceed by the following steps:

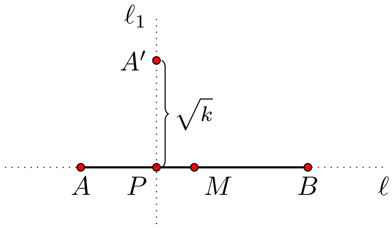


Step 1

Step 2

Step 1: Use (HH2) to construct the fold f_1 that is the perpendicular bisector of AB. Denote by M the midpoint of AB.

Step 2: Using (HH5), construct the fold f_2 to place A to the point A' onto the line ℓ_1, that is perpendicular to AB at P. Observe that the segments AM, $A'M$ and BM have the same length. Hence the triangle $\triangle AA'B$ is inscribed into the circle with the centre in M and radius AM.



Therefore, the length of $A'P$ is $\sqrt{k}$ by Remark 5.5. ■

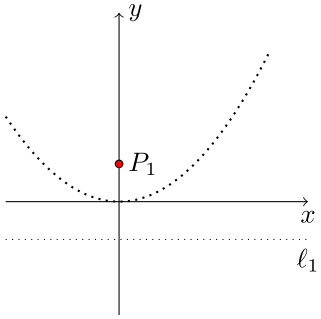
7.11. **Corollary.** *The field* $\mathbb{O}$ *extends the field* $\mathbb{E}$.

We will now focus on showing that $\mathbb{E} \neq \mathbb{O}$ holds. In fact, we will see that $\mathbb{O}$ is closed under *cubic roots* in Corollary 7.14. First, we need to understand cubic equations and their relationship to axiom (HH6).

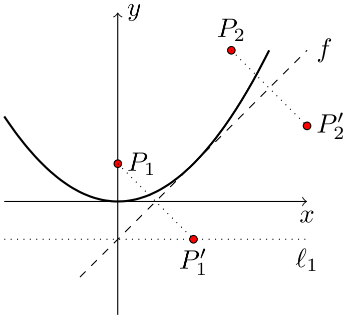
7.12. **Remark (Axiom (HH6) and cubic equations).** It can be shown that axiom (HH6) is equivalent to solving certain cubic equations. To that end, let P_1 be the point with coordinates $(0, 1)$, let ℓ_1 be the line $y = -1$. The parabola with focus P_1 and directrix ℓ_1 has, by Remark 7.3, the equation

$$y = \frac{1}{4}x^2$$

See the picture



Suppose that P_1 is folded to the point P_1' with coordinates $(t, -1)$ and suppose P_2 has coordinates (a, b). We want to determine the coordinates (x, y) of the point P_2'.



Since the equation of f is

$$y = \frac{t}{2}x - \frac{t^2}{4}$$

and since f is the perpendicular bisector of P_2P_2', we must have

$$\frac{t}{2} = -\frac{x-a}{y-b}$$

Moreover, f passes through the midpoint of the segment P_2P_2', hence the equality

$$\frac{y+b}{2} = \frac{t}{2} \cdot \frac{a+x}{2} - \frac{t^2}{4}$$

holds. Substituting the above expression for $t/2$, we obtain

$$\frac{y+b}{2} = -\frac{x-a}{y-b} \cdot \frac{a+x}{2} - \Big(\frac{x-a}{y-b}\Big)^2$$

In other words, the equality

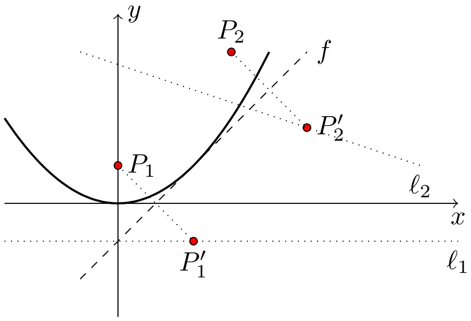
$$(y+b) \cdot (y-b)^2 = -(x^2 - a^2) \cdot (y-b) - 2 \cdot (x-a)^2$$

or, equivalently, the equality

$$x^2(-b+2) + x(-2a) + y^3 + y^2(-b) + y(-a^2 - b^2) + x^2y + (a^2b + b^3 + a^2) = 0$$

holds.

Suppose now that ℓ_2 is given by the equation $Ax + By + C = 0$. Axiom (HH6) requires us to put P_2' on the line ℓ_2:



Hence the coordinates (x, y) of the point P_2' satisfy the system

$$x^2(-b+2) + x(-2a) + y^3 + y^2(-b) + y(-a^2 - b^2) + x^2y + (a^2b + b^3 + a^2) = 0 \qquad Ax + By + C = 0$$

Since $A^2 + B^2 \neq 0$, we can express either y or x from the equality $Ax + By + C = 0$, substitute it to the first equality and obtain a cubic equation either in y or in x. Therefore, by using axiom (HH6), we solve a certain cubic equation.

Remark 7.12 shows that (HH6) is equivalent to solving a *certain* cubic equation. We show now that *any* cubic equation with coefficients in $\mathbb{O}$ can be solved by origami axioms. Let us recall first that *any* cubic polynomial $p(x)$ in $\mathbb{R}[x]$ must have at least one root in $\mathbb{R}$. This is seen as follows:

(1) Since $p(x)$ is of odd degree, either

$$\lim_{x\to-\infty} p(x) = -\infty \qquad \text{and} \qquad \lim_{x\to+\infty} p(x) = +\infty$$

or

$$\lim_{x\to-\infty} p(x) = +\infty \qquad \text{and} \qquad \lim_{x\to+\infty} p(x) = -\infty$$

hold.

(2) From (1) there exists a closed interval $[a; b]$ such that $p(a) \cdot p(b) < 0$. Since $x \mapsto p(x)$ is a continuous function, then, by Intermediate Value Theorem[23] of calculus, there exists c in $(a; b)$ such that $p(c) = 0$.

7.13. **Task for you (Solving cubic equations).** *A real root of a cubic polynomial in $\mathbb{O}[x]$ can be origami-constructed.*

Hints for the proof. Without loss of generality we may assume that our cubic equation is normalised, i.e., that we are to solve the equation

$$x^3 + ax^2 + bx + c = 0$$

with a, b, c in $\mathbb{O}$. Let us introduce the substitution

$$X = x - \frac{a}{3}$$

Then we are to solve the equation

$$X^3 + \underbrace{\frac{3b - a^2}{3}}_{=A} X + \underbrace{\frac{2a^3 - 9ab + 27c}{27}}_{=B} = 0$$

where A and B are in $\mathbb{O}$. Thus, we will show how to find a real root of any cubic equation

$$x^3 + Ax + B = 0$$

where A, B are in $\mathbb{O}$ and, without loss of generality, we can assume $B > 0$.[24]

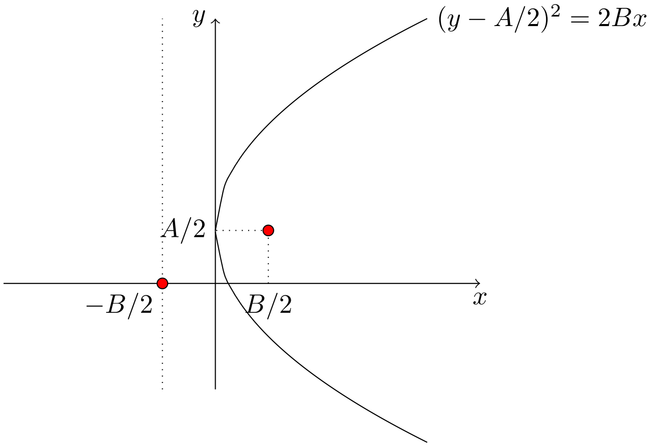
Let us consider two parabolas

$$(y - \frac{1}{2}A)^2 = 2Bx \qquad \text{and} \qquad y = \frac{1}{2}x^2$$

[23] The theorem states: if $f : [a; b] \longrightarrow \mathbb{R}$ is a continuous function and $f(a) \cdot f(b) < 0$, then there exists c in $(a; b)$ such that $f(c) = 0$. See, e.g., [31].

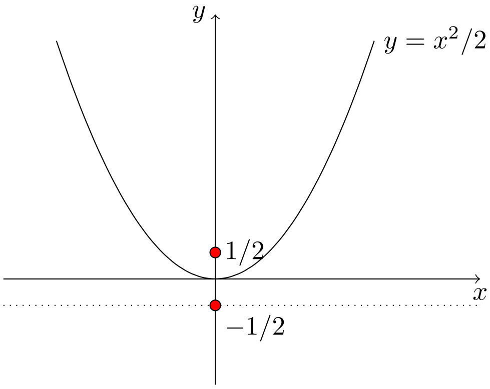
[24] If $B = 0$, solving $x^3 + Ax + B = 0$ is easy. If $B < 0$, make the substitution $X = -x$, and then consider the equation $X^3 + AX - B = 0$.

The parabola on the left has the graph



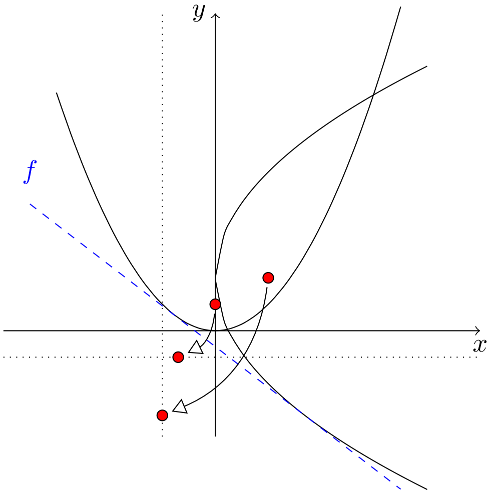
having the focus in the point $(B/2, A/2)$ and directrix $x = -B/2$.

The parabola on the right has the graph



having the focus in the point $(0, 1/2)$ and directrix $y = -1/2$.

We use axiom (HH6), folding $(B/2, A/2)$ to $x = -B/2$ and $(0, 1/2)$ to $y = -1/2$:



The fold f is the line of the form $y = px + q$ and it is the common tangent to both parabolas. We claim that p is the desired solution, i.e., we claim that

$$p^3 + Ap + B = 0$$

holds.

Let us denote by (x_1, y_1) and (x_2, y_2) the tangent points on $px + q$ to $(y - A/2)^2 = 2Bx$ and $y = x^2/2$, respectively. By taking the derivative of $(y - A/2)^2 = 2Bx$ implicitly, we conclude that $2(y - A/2)y' = 2B$. By taking the derivative of $y = x^2/2$, we obtain $y' = 2x$. Hence, at (x_1, y_1), we obtain $2(y_1 - A/2)p = 2B$.

Thus $y_1 = B/p + A/2$ and, consequently, $x_1 = B/(2p^2)$. Analogously, at (x_2, y_2), we have $x_2 = p$ and $y_2 = p^2/2$.

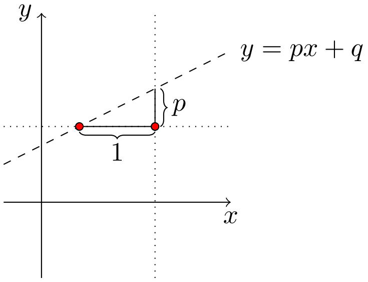
Therefore

$$p = \frac{y_2 - y_1}{x_2 - x_1} = \frac{p^2/2 - (B/p + A/2)}{p - B/(2p^2)}$$

which gives, after simplification, the claimed equality

$$p^3 + Ap + B = 0$$

To finalise the proof, we have to construct the slope p of the line $y = px + q$. This is done by choosing a point on the line, another point at the distance 1 in the direction of axis x, and by erecting a perpendicular line through that point.



Axiom (HH4) suffices to do that. See Task 7.6. ■

7.14. Corollary (Closedness of $\mathbb{O}$ under cubic roots). *Suppose r is in $\mathbb{O}$. Then $\sqrt[3]{r}$ is in $\mathbb{O}$.*

Hints for the proof. Without loss of generality, suppose that $r > 0$. Then the simultaneous tangent f to $y^2 = 2rx$ and $y = x^2/2$ has the slope $\sqrt[3]{r}$. ■

7.15. Task for you (Characterisation of $\mathbb{O}$). *For a real number r, the following conditions are equivalent:*

(1) $r \in \mathbb{O}$.

(2) *There exists a sequence $\mathbb{F}_0 \preceq \mathbb{F}_1 \preceq \ldots \preceq \mathbb{F}_N$ of fields such that $\mathbb{F}_0 = \mathbb{Q}$, $r \in \mathbb{F}_N$, and $[\mathbb{F}_{i+1} : \mathbb{F}_i] = 2$ or $[\mathbb{F}_{i+1} : \mathbb{F}_i] = 3$, for all $i = 0, \ldots, N-1$.*

Hints for the proof. (1) implies (2). An origami-constructible number is given by repeating axioms (HH1)–(HH7) finitely many times. Since Tasks 7.10 and 7.13 show that we are dealing with field extensions of degree up to 3, the assertion follows.

(2) implies (1). We proceed by induction on $N \geq 0$.

(a) If $N = 0$, then r is a rational number, which is origami-constructible.

(b) Suppose $N > 0$ and suppose that any element of $\mathbb{F}_{N-1}$ is origami-constructible. Since either $\mathbb{F}_N = \mathbb{F}_{N-1}(\sqrt{k})$ or $\mathbb{F}_N = \mathbb{F}_{N-1}(\sqrt[3]{k})$ for some k in $\mathbb{F}_{N-1}$, the elements of $\mathbb{F}_N$ are origami-constructible by Tasks 7.10 and 7.13.

■

7.16. Corollary (The origami constructible numbers). *If r is an origami-constructible number, then the equality $[\mathbb{Q}(r) : \mathbb{Q}] = 2^A \cdot 3^B$ holds for some natural numbers A and B.*

Hints for the proof. By Task 7.15 there exists a sequence $\mathbb{F}_0 \preceq \mathbb{F}_1 \preceq \ldots \preceq \mathbb{F}_N$ of fields such that $\mathbb{F}_0 = \mathbb{Q}$, $r \in \mathbb{F}_N$, and $[\mathbb{F}_{i+1} : \mathbb{F}_i] = 2$ or $[\mathbb{F}_{i+1} : \mathbb{F}_i] = 3$, for all $i = 0, \ldots, N-1$. Hence

$$[\mathbb{F}_N : \mathbb{Q}] = 2^a \cdot 3^b, \quad \text{where } a + b = N$$

by Task 2.12. By the same task, the degree $[\mathbb{Q}(r) : \mathbb{Q}]$ divides $[\mathbb{F}_N : \mathbb{Q}]$, since $\mathbb{Q} \preceq \mathbb{Q}(r) \preceq \mathbb{F}_N$ holds. Thus, there exist natural numbers A, B such that $[\mathbb{Q}(r) : \mathbb{Q}] = 2^A \cdot 3^B$ holds. ■

7.17. **Remark (Comparing straightedge and compass constructions and origami constructions).** Corollaries 5.12 and 7.16 tell us the precise difference between real numbers that are constructible by straightedge and compass and real numbers that are origami-constructible. For the record

(1) r is in $\mathbb{E}$ iff $[\mathbb{Q}(r) : \mathbb{Q}] = 2^A$ for some natural number A.
(2) r is in $\mathbb{O}$ iff $[\mathbb{Q}(r) : \mathbb{Q}] = 2^A \cdot 3^B$ for some natural numbers A and B.

We will use this difference in the next chapter to tackle the Ancient Greek problems by origami techniques.

8. Cracking the Ancient Greek Problems with origami

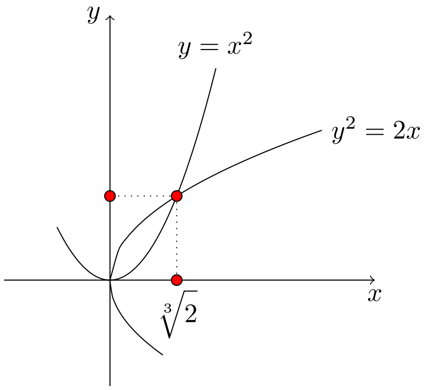
We know by Corollary 7.16 that origami axioms yield more powerful constructions than those by compass and straightedge. In fact, we show now that origami axioms allow us to *double the cube* and to *trisect an angle*. By Corollary 7.16 it is, however, still impossible to square the circle using origami: every origami-constructible number is algebraic over $\mathbb{Q}$ by Corollary 7.16, but π is transcendental over $\mathbb{Q}$ by Corollary 4.7.

Doubling the cube. The problem of doubling the cube reduces to the construction of $\sqrt[3]{2}$. This number can be easily constructed as the intersection of two parabolas. Indeed, the solution of $x^3 - 2 = 0$ can be found by the system of equations

$$y = x^2 \qquad y^2 = 2x$$

by Task 7.13.

is given by solving the equation $x^4 = 2x$, or, equivalently, by solving $x \cdot (x^3 - 2) = 0$. Therefore $x = 0$ or $x = \sqrt[3]{2}$. Geometrically, we are to compute the intersection of two parabolas:

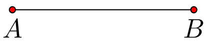


How do we construct the intersection using origami axioms? There are several possible ways. The next solution is due to Peter Messer from his paper [32] from 1986. The construction has two steps:

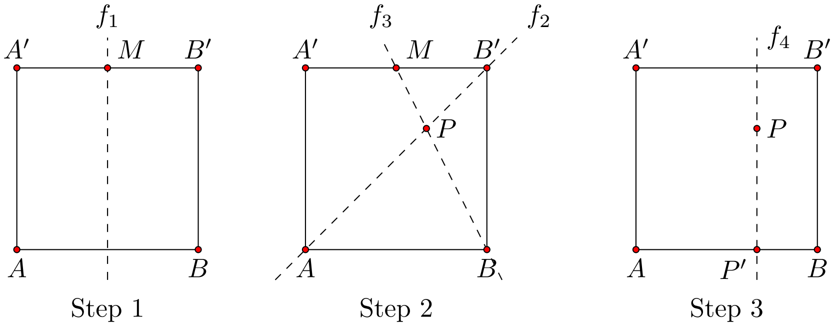
(I) We first divide a segment into thirds.
(II) We use (I) to construct $\sqrt[3]{2}$.

The individual steps are performed as follows:

(I) Consider the segment



and proceed by the following steps:[25]



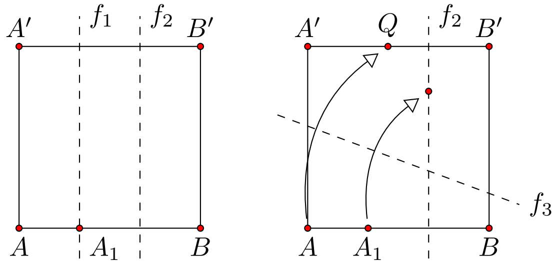
Step 1: Construct the square $ABB'A'$, using (HH4) repeatedly. Then use (HH3) to find the fold f_1 that bisects the parallel lines AA', BB'. Denote by M the midpoint of $A'B'$, given by f_1.

Step 2: Construct the folds f_2 and f_3, using (HH1). Denote by P the intersection of f_2 and f_3.

Step 3: Use (HH4) to create the fold f_4. Denote by P' the intersection of AB and f_4.

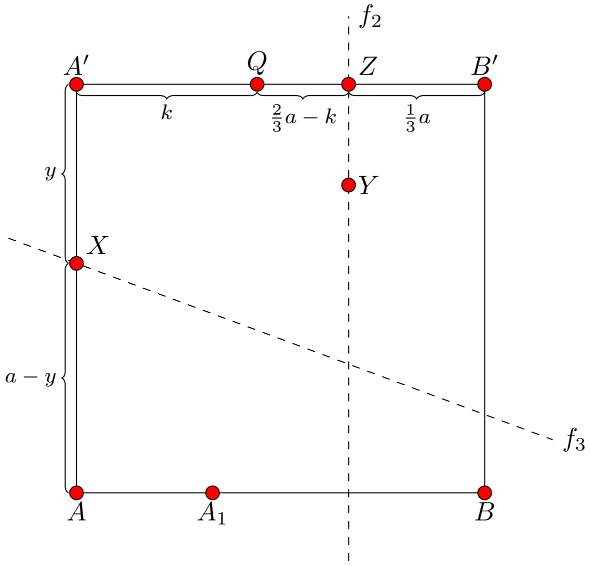
We claim that the length of $P'B$ is one-third of the segment AB. Indeed, putting the origin in the point A and putting the edge of the square equal to a, the point P is the intersection of lines $y = x$ (fold f_2) and $y = -2(x - a)$ (fold f_3), i.e., the coordinates of P are $((2/3)a, (2/3)a)$.

(II) Suppose we have a square $ABB'A'$ with AB and $A'B'$ divided into thirds by construction (I) above. Let f_1 and f_2 be the folds giving the trisection of AB, as indicated below. Denote by A_1 the intersection of f_1 and AB.



Then use (HH6) to put A on $A'B'$ as the point Q and to put A_1 on the line f_2, as indicated on the right above. We claim that the ratio of QB' and $A'Q$ is $\sqrt[3]{2}$.

Suppose the edge of the square has length a. Denote the various lengths as follows:



[25] Ths method was developed by Kazuo Haga, see [26].

We want to prove that the equality

$$\frac{a-k}{k} = \sqrt[3]{2}$$

holds. We proceed in three steps:

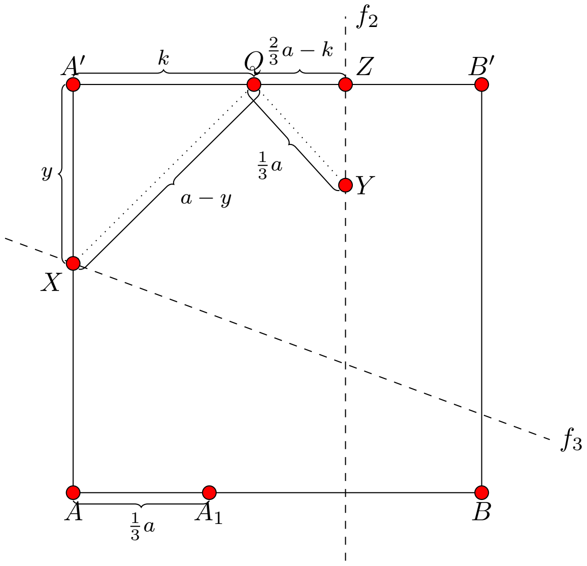
(a) Since f_3 is the perpendicular bisector of AQ, we have $|XQ| = a - y$. Hence, using Pythagoras' Theorem on the right-angled triangle $\triangle XA'Q$, we obtain the equality

$$(a-y)^2 = k^2 + y^2$$

that yields, after simplification, the equality

$$y = \frac{a^2 - k^2}{2a}$$

(b) We claim next that the right-angled triangles $\triangle XA'Q$ and $\triangle QZY$ are similar: this follows from the fact that f_3 is the perpendicular bisector of both AQ and A_1Y. From that it also follows that $|QY| = |AA_1| = (1/3)a$.



Thus, we have the equality

$$\frac{y}{a-y} = \frac{(2/3)a - k}{(1/3)a}$$

(c) Substituting the equality $y = (a^2 - k^2)/(2a)$ from (a) into

$$\frac{y}{a-y} = \frac{(2/3)a - k}{(1/3)a}$$

and simplifying yields

$$\frac{a^2 - k^2}{a^2 + k^2} = \frac{2a - 3k}{a}$$

or, equivalently, we have the equality

$$a^3 - ak^2 = \underbrace{2a^3 - 3ka^2 + 2k^2a - 3k^3}_{=(a-k)^3 + a^3 - ak^2 - 2k^3}$$

That is, we have the equality

$$2k^3 = (a-k)^3$$

or, equivalently, the equality

$$\frac{a-k}{k} = \sqrt[3]{2}$$

as desired.

Trisecting an angle. The problem of trisecting a general angle α reduces to constructing a root of a certain cubic equation. Indeed, since the equation

$$\cos 3\alpha = 4\cos^3\alpha - 3\cos\alpha$$

holds for any α, we can construct $\cos\alpha$ by solving the cubic equation

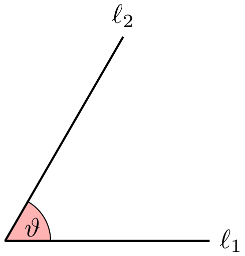
$$4x^3 - 3x - \cos 3\alpha = 0$$

which can be done, for example, by Task 7.13. Having constructed $\cos\alpha$, it is easy to construct the angle α.

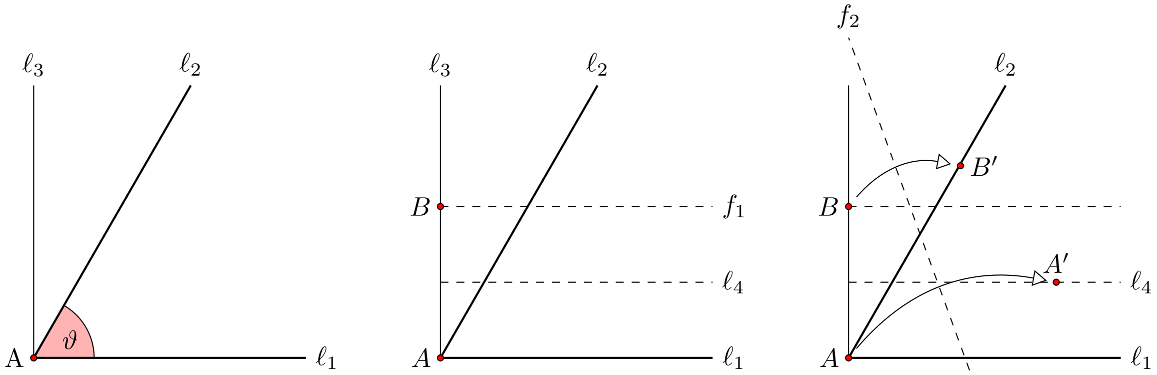
There is a neat way how to trisect an angle using the origami axioms. The following construction is due to Hisahi Abe, see [12].

(1) We will first present the trisection of an *acute* angle, i.e., an angle between 0 and $\pi/2$.

Suppose we have lines ℓ_1 and ℓ_2 determining an acute angle ϑ that we want to trisect:



We proceed by the following steps:



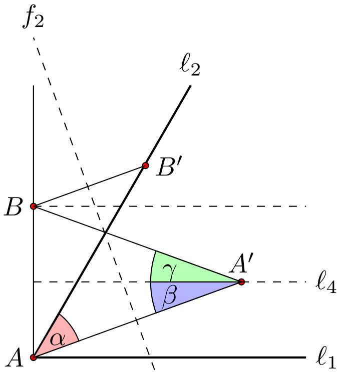
Step 1 Step 2 Step 3

Step 1: Denote by A the point at the vertex of the angle that we want to trisect. Use (HH4) to construct the line ℓ_3, perpendicular to ℓ_1 at point A.

Step 2: Choose any point B on ℓ_3, on the same side of ℓ_1 as the angle ϑ. Construct the fold f_1 by (HH4) as the parallel to ℓ_1 passing through B and then denote by ℓ_4 the bisector of f_1 and ℓ_1, constructed by (HH2).

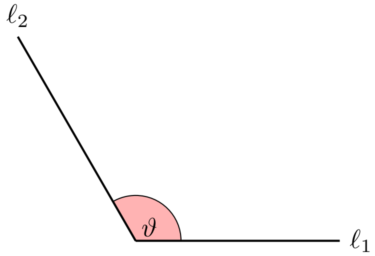
Step 3: Use (HH6) to obtain the fold f_2 putting A to ℓ_3 and B to ℓ_2. Denote by A' and B' the resulting points, respectively.

We claim that the line AA', together with ℓ_1, determines the angle $\vartheta/3$. Indeed, consider the picture

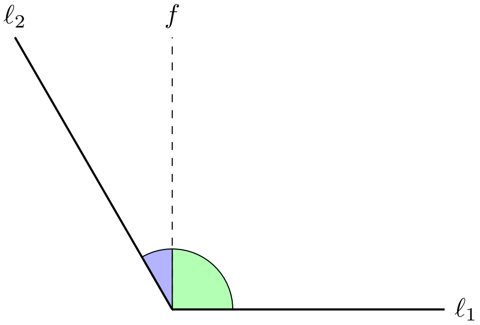


Since f_2 is the bisector of AA', we have $\alpha = \beta + \gamma$. Since ℓ_4 is the bisector of the angle $\measuredangle BA'A$, we have $\beta = \gamma$. Hence $\alpha = 2\beta$. Since $\vartheta = \alpha + \beta$, we have $\beta = \vartheta/3$. But β is precisely the angle between AA' and ℓ_1, since the lines ℓ_1 and ℓ_4 are parallel.

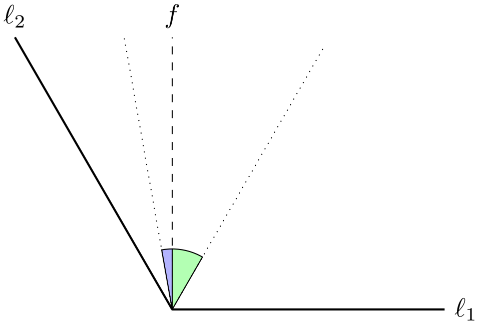
(2) Suppose the angle ϑ is *obtuse*, i.e., suppose that ϑ is between $\pi/2$ and π.



Proceed by the following steps:



Step 1



Step 2

Step 1: Use (HH4) to construct the fold f that is perpendicular to ℓ_1 and that divides the angle ϑ into an acute angle and the right angle.

Step 2: Trisect the acute angle using part (1) above. Then trisect the right angle by constructing the angle $\pi/6$. The sum of these angles is the trisection of the angle ϑ.

9. A bit of Galois Theory and final remarks

Extending fields with the prospect of obtaining roots of polynomials was systematically treated by the French mathematician *Évariste Galois* at the beginning of the 19th century. Before we say a bit about Galois' colourful (if short) life, we indicate the main ideas of the theory, called *Galois Theory* nowadays.

Galois Theory in a nutshell. The idea of Galois was as follows: start with a polynomial $p(x)$ in $\mathbb{F}[x]$. Then one can construct a certain extension $\mathbb{F} \preceq \mathbb{F}'$ such that $\mathbb{F}'$ is the smallest field over which $p(x)$ can be split into linear factors. This extension $\mathbb{F}'$ is called the *splitting field* of $p(x)$. It can be proved that the splitting field $\mathbb{F}'$ of $p(x)$ is essentially unique and that the roots of $p(x)$ show a symmetry that can be captured by certain groups.

More in detail: the above extension $\mathbb{F} \preceq \mathbb{F}'$ is called a *Galois extension*, if the degree $[\mathbb{F}' : \mathbb{F}]$ of the extension is equal to the number of elements of the group $\mathrm{Aut}(\mathbb{F}'/\mathbb{F})$ of field isomorphisms $f : \mathbb{F}' \longrightarrow \mathbb{F}'$ such that $f(r) = r$ for all r in $\mathbb{F}$.

The *Fundamental Theorem of Galois Theory* then reads as follows:

> *Under certain additional hypotheses on $\mathbb{F} \preceq \mathbb{F}'$ there is a bijective correspondence between subgroups of $\mathrm{Aut}(\mathbb{F}'/\mathbb{F})$ and subfields of $\mathbb{F}'$ containing $\mathbb{F}$. Moreover, the correspondence reverses the order,*[26] *i.e., if the group G_1 corresponds to the field $\mathbb{F}_1$ and the group G_2 corresponds to the field $\mathbb{F}_2$, then*
>
> $$G_1 \text{ is a subgroup of } G_2 \quad \text{iff} \quad \mathbb{F}_2 \text{ is a subfield of } \mathbb{F}_1$$
>
> *holds.*

[26] Such a correspondence is often called a *Galois connection*. It is an instance of the general notion of *adjunction* from Category Theory. See, e.g., [42].

As a corollary of this result, one can derive that a polynomial equation $p(x) = 0$ is *solvable in radicals*, if its Galois group $\mathrm{Aut}(\mathbb{F}'/\mathbb{F})$ has a certain property, pertaining to its subgroups.

Hence, for example, the famous result of *Niels Henrik Abel* [1]:[27]

> *For every $n \geq 5$, a general polynomial equation $p(x) = 0$ of degree n with rational coefficients is not solvable in radicals.*

is a corollary of the general theory. Galois Theory, however, allows for a finer analysis of roots of polynomial equations.

A brief biography of Évariste Galois. Évariste Galois was born on 25 October 1811 in Bourg La Reine (Paris' outskirts) in France. Until he was 12 years old, Galois' only teacher was his mother, who taught him Greek, Latin and religion. Galois started his formal education on 6 October 1823 at the Lycée of Louis-le-Grand. France suffered turmoils of revolution at that time. Even in Galois' first term at the lyceum, there was a minor rebellion and forty students were expelled because of that. Galois was not involved in the rebellion and he proceeded with his studies. The turning point for him was February 1827, when Galois enrolled his first mathematical classes. Mathematics became his passion and Galois' school reports began to describe him as *singular*, *bizarre*, *original* and *closed*.

In 1828 Galois took the examination at the leading university in France, the École Polytechnique in Paris. He obviously wanted to pursue the academic career, but he failed the exam. Galois did not give up: in April 1829 his first mathematics paper was published [13] but he failed at the École Polytechnique exam for the second time. Galois had to take the Baccalaureate examinations[28] and then he entered the École Normale.

In 1830, the political mess in France went beyond the usual. Charles X left France and the streets were rioting again. While the students of the Polytechnique were making history in the streets, the students of the École Normale were locked in by the school's director. In December 1830, Galois wrote an article in the Gazette des Écoles, attacking the director for locking the students into the school. Galois was expelled and he joined the Artillery of the National Guard, a Republican branch of the militia. On 31 December 1830 the Artillery of the National Guard was abolished by Royal Decree since the new King Louis-Phillipe d'Orleans perceived it as a threat to the throne. During one of the meetings celebrating the National Guard Galois proposed a toast to King Louis-Philippe with a dagger above his cup and he was arrested the next day, and acquitted in June 1831.

On the Bastille Day, 14 July 1831, Galois was arrested for wearing the uniform of the Artillery of the National Guard, loaded gun, several pistols, and dagger. He was sentenced to prison again, this time for a nine months sentence. When cholera broke out in March 1832 in Paris, the prisoners, including Galois, were transferred to Faultrier clinic. There, Galois fell in love with Stéphanie-Felice Poterine-Dumotel, the daughter of a local physician. After being released in April, Galois started exchanging letters with Stéphanie. From the tone of her replies though, it is clear that Stéphanie wanted to end the affair.

In the spring of 1832, the most militant part of the republicans formed the ominously named *la Société des amis du peuple*. The group was ready to stage any riot and violence to overthrow the reign of King Louis-Phillipe d'Orleans. Needless to say, Galois joined the society and became its active member as soon as he could.

On 30 May 1832 Galois fought a duel with Perscheux d'Herbinville, the 22 years-old member of the Artillery of the National Guard. The reason for the duel, although not quite clear, was certainly linked with Stéphanie.[29] During the night preceeding the duel, Galois tried to finish a manuscript of a paper. In the

[27] Niels Henrik Abel (5 August 1802, Nedstrand, Denmark-Norway – 6 April 1829, Froland, Norway) was a Norwegian mathematician. He developed, independently of Galois, group theory and he proved that general polynomial equations of degree ≥ 5 cannot be solved in radicals. Abel's outstanding results in field theory led to establishing *Abel Prize* in 1899 as a "Nobel Prize" in mathematics.

[28] Galois passed in maths with flying colours but he did not impress his examiner in literature, who reported: *This is the only student who has answered me poorly, he knows absolutely nothing. I was told that this student has an extraordinary capacity for mathematics. This astonishes me greatly, for, after his examination, I believed him to have but little intelligence.*

[29] This version is supported by the memoirs of Alexandre Dumas, see [7], and by Galois' own words that he was "victime d'une infâme coquette". A different account of Galois' death is given by Laura Toti Rigatelli in her book [39]. According to this book, the duel was staged by Galois and his militant comrades. Galois' intention was to be killed in the duel to produce a corpse that would stir a further uprising of republican revolutionaries. If that theory is true, Galois failed tragically: although

margin of the manuscript there is a note saying: *There is something to complete in this demonstration. I do not have the time.* This note, perhaps, led to the legend that Galois spent the night before his duel writing all he knew about group theory.

Galois was wounded in the duel and was left abandoned both by d'Herbinville and by his own seconds. After some time a peasant found wounded Galois. Galois died of peritonitis in Cochin hospital at 10 in the morning of 31 May 1832, aged only 22 years.

During his lifetime, Galois published six papers. After his death, Galois' brother Alfred and Galois' friend Auguste Chevalier copied Galois' mathematical papers and sent them to various mathematicians, including Karl Friedrich Gauss and Carl Gustav Jacob Jacobi. Galois wished to know Gauss' and Jacobi's opinions on his work. We know of no comments of Gauss and Jacobi on Galois' work. After Galois' work reached Joseph Liouville, the results were presented to the French Academy. Liouville published Galois' papers in 1846 in the survey paper [29].

Comments on the literature.

(1) **Constructions with straightedge and compass.** The best starting point for classical constructions of geometry is *Elements* [11] by Euclid. Two books [18] and [19] by Sir Thomas Little Heath give, unsurpassed in their scope and detail, an overview of the classical Greek geometry and mathematics, ranging from the earliest work of Thales of Miletus till Diophantus of Alexandria.

Beautiful proofs involving triangles, polygons, etc, can be found in the book [4] by Claudi Alsina and Roger B. Nelsen. Another collection of beautiful proofs, including Hippocrates' squaring of the lunes, is the book [10] by William Dunham. A precise and detailled account of squaring the lunes is given in Chapter 16 of the book [38] (in Russian) by Mikhail Mikhailovich Postnikov.

The classical straightedge & compass constructions are treated in the book [16] by Charles Robert Hadlock. Hadlock's book is also a very nice and gentle introduction to Galois Theory.

A bit more advanced branches of geometry are covered in [17] by Robin Hartshorne. Basic results on field extensions can also be found there.

(2) **Origami constructions.** The Master's Thesis [27] by Hwa Young Lee covers all the basic information about origami constructions. More detailed and advanced treatment of origami-constructible numbers is in the paper [3] by Roger C. Alperin and Robert J. Lang. The papers [2] by Roger C. Alperin and [26] by Robert J. Lang give a very readable overview of subfields of the reals that are related to various geometric constructions.

Chapter 10 of the book [8] by David A. Cox gives a nice overview of how geometric constructions (both origami and straightedge & compass) sit within the realm of Galois Thoery.

Thomas Hull's book [21] is a collection of projects in origami, leading to an understanding of mathematical principles behind paper folding. See http://mars.wne.edu/~thull/projectorigami/toc.html for more information.

(3) **Abstract algebra and Galois Theory.** Galois Theory is an important part of the branch of mathematics often called *Abstract Algebra*, i.e., of the study of rings, fields, etc. There are very many great books on Abstract Algebra. Let us mention just a few of them.

Perhaps the first book that appeared on the topic, covering all the essentials, is Bartel Leendert Van der Waerden's book [41]. The first edition is from the 1930's, but new editions largely expand the original text.

We already mentioned Charles Robert Hadlock's book [16] as a gentle introduction to Galois Theory. A bit more advanced, but still great for self-study, is the book [5] by Juliusz Brzeziński.

To cite just two more advanced books: David A. Cox's book [8] is a thorough book on Galois' Theory and so is the excellent book [38] (in Russian) by Mikhail Mikhailovich Postnikov.

(4) **Galois' life and revolutions in France.** Perhaps the best-known account of Galois' life is the book [22] by the Polish physicist Leopold Infeld. Another account is given by the Italian historian of mathematics Laura Toti Rigatelli in [39]. The information about Galois' adversary is given in the papers [7] by Olivier Courcelle. It is very likely, though, that we might never know all the details about Galois' duel and his death.

A very detailed survey of the French Revolution is on the nearly 1 000 pages of Simon Schama's book [40]. The book covers the period from roughly August 1776 to July 1794. The revolutionary era of France in the 1830's, connected to Galois' life, is the topic of the book [34] by David H. Pinkney.

about 3 000 people, ready to attack the police, gathered at Galois' funeral, the crowd dispersed itself when rumours about a "more important corpse" for the revolutionary cause spread around.

References

[1] N. H. Abel, Beweis der Unmöglichkeit, algebraische Gleichungen von höheren Graden als dem vierten allgemein aufzulösen, *J. Reine Angew. Math.* 1 (1826), 65–84. 53

[2] R. C. Alperin, A mathematical theory of origami constructions and numbers, *New York J. Math.* 6 (2000), 119–133. 54

[3] R. C. Alperin and R. J. Lang, One-, two-, and multi-fold origami axioms, In: *Origami*^{4}, A. K. Peters, Natick, MA, 2009, 371–393. 37, 54

[4] C. Alsina and R. B. Nelsen, *Charming proofs: A journey into elegant mathematics*, The mathematical Association of America, 2010. 54

[5] J. Brzeziński, *Galois Theory through exercises*, Springer, 2018. 54

[6] G. Cantor, Über eine Eigenschaft des Inbegriffs aller reellen algebraischen Zahlen, *J. Reine Angew. Math.* 77 (1874), 258–262. 18

[7] O. Courcelle, L'adversaire de Galois (I) and L'adversaire de Galois (II), CNRS, 2015. 53, 54

[8] D. A. Cox, *Galois theory*, John Wiley & Sons, 2012. 54

[9] D. S. Dummit and R. M. Foote, *Abstract algebra*, John Wiley & Sons, 2004. 12

[10] W. Dunham, *Journey through genius: The great theorems of mathematics*, Penguin, 1991. 54

[11] Euclid, *Elements*, 300BC. 1, 2, 23, 32, 54

[12] K. Fusimi, Trisection of angle by Abe, *Saiensu Supplement* 8 (1980). 51

[13] É. Galois, Demonstration d'un théorème sur les fractions continues périodiques, *Annales de mathématiques pures et appliquées* XIX (1829), 294–301. 53

[14] G. Glaeser, H. Stachel and B. Odehnal, *The universe of conics*, Springer, 2016. 37, 40

[15] W. H. Greub, *Linear algebra*, Springer, 1967. 40

[16] C. R. Hadlock, *Field theory and its classical problems*, Mathematical Association of America, 2014. 54

[17] R. Hartshorne, *Geometry: Euclid and beyond*, Springer, 2000. 54

[18] T. L. Heath, *A history of Greek mathematics, Vol. 1: From Thales to Euclid*, Cambridge University Press, 2014. 54

[19] T. L. Heath, *A history of Greek mathematics, Vol. 2: From Aristarchus to Diophantus*, Cambridge University Press, 2014. 54

[20] C. Hermite, Sur la fonction exponetielle, *C. R. Math. Acad. Sci. Paris* 77 (1873), 18–24. 18

[21] T. Hull, *Project Origami: Activities for exploring mathematics*, CRC Press, 2012. 54

[22] L. Infeld, *Whom the gods love: The story of Évariste Galois*, National Council of Teachers, 1978. 54

[23] N. Jacobson, *Basic algebra* Vols. 1 and 2, W. H. Freeman, 1980. 12

[24] J. Justin, Résolution par le pliage de l'équation du troisieme degré et applications géométriques, In: *Proceedings of the first international meeting of origami science and technology*, Ferrara, Italy,1989, 251–261. 36

[25] K. Kendig, *Conics*, The Mathematical Association of America, 2005. 37

[26] R. J. Lang, *Origami and geometric constructions*, 2015. 49, 54

[27] H. Y. Lee, *Origami-constructible numbers*, MA Thesis, University of Athens, Georgia USA, 2017. 54

[28] F. Lindemann, Über die Zahl π, *Math. Ann.* 20 (1882), 213–225. 18

[29] J. Liouville (ed), Œuvres mathématiques d'Évariste Galois, *Annales de mathématiques pures et appliquées* XI (1846), 381–444. 54

[30] J. Liouville, Sur les classes très étendues de quantités dont la valeur nest ni algébrique ni même réductible à des irrationelles algébriques, *J. Math. Pures Appl.* 16 (1851), 133–142. 18

[31] J. E. Mardsen, *Elementary classical analysis*, W. H. Freeman & co., 1993. 45

[32] P. Messer, Problem 1054, *Crux Mathematicorum* 12 (1986), 284–285. 48

[33] I. Niven, *Irrational numbers*, John Wiley & Sons, 1956. 19

[34] D. H. Pinkney, *French Revolution of 1830*, Princeton University Press, 2019. 54

[35] Plato, *Meno*, (translated by C. Woods), SSRN, 2012. 28

[36] Plato, *Republic*, 27

[37] M. M. Postnikov and A. Schenitzer, The problem of squarable lunes, *Amer. Math. Monthly* 107 (2000), 645–651. 33, 34

[38] M. M. Постников, *Теория Галуа* (M. M. Postnikov, *Galois theory*), Faktorial Press, Moscow, 2003. 33, 34, 54

[39] L. T. Rigatelli, *Évariste Galois: 1811 – 1832*, Birkhäuser, 1996. 53, 54

[40] S. Schama, *Citizens: A chronicle of the French revolution*, Penguin, 2004. 54

[41] B. L. Van der Waerden, *Algebra*, Springer, 1991. 54

[42] J. Velebil, *Categorical methods in universal algebra*, TACL 2017 Summer School Lecture Notes, 22 June 2017. 52

[43] J. Velebil, *Dedekind's construction of the reals*, manuscript, FEL ČVUT, Praha, 2019. 2, 3, 10

[44] L. Wantzel, Recherches sur les moyens de reconnaître si un Problème de Géométrie peut se résoudre avec la règle et le compas, *J. Math. Pures Appl.* 2 (1837), 366–372. 28

[45] K. Weierstrass, Zu Lindemann's Abhandlung "Über die Ludolphsche Zahl", *Sitzungsber. Preuss. Akad. Wiss. Berlin* 1885, 1067–1085 18

Department of Mathematics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

E-mail: velebil@math.feld.cvut.cz