A Course on ALGEBRA

by

Ahmet K. Feyzioğlu

II. CİLT

A Course on

ALGEBRA

by

Ahmet K. Feyzioğlu

II. CİLT

Dedicated to the memory of my father

SABAHATTİN FEYZİOĞLU

Cataloging - in - publication Data

Feyzioğlu, Ahmet K.

A course on algebra
Bibliographyip

1. Algebra 1. Title
512

ISBN: 975-518-014-1

Boğaziçi University Publication No:496
printed in Turkey
Boğaziçi University Printing Office
Bebek, İstanbul 80815 Turkey

Table of Contents

CHAPTER 4

Vector Spaces

§39
Definition and Examples

The term "vector" is familiar to the reader from Physics. Such physical magnitudes as displacement, force, torque, momentum etc. are vectors. Vectors can be added (by the parallelogram law) and multiplied by real numbers, which are called scalars in this context. In this chapter, we introduce systems of objects which can be added and multiplied by scalars.

39.1 Definition: Let K be a field and let $(V,+)$ be an (additively written) abelian group. Suppose that, to each pair (α,v) in $K \times V$, there corresponds a unique element of V, denoted by $\alpha.v$, such that the following equations hold for all $\alpha,\beta \in K$, $v,w \in V$:

(1) $\alpha.(v+w) = \alpha.v + \alpha.w$

(2) $(\alpha+\beta).v = \alpha.v + \beta.w$

(3) $\alpha.(\beta.v) = (\alpha\beta).v$

(4) $1.v = v$ (1 is the identity of K).

In this case, the ordered quadruple $(V,+,K,\cdot)$ is called a *vector space over* K, or a *K-vector space*. The elements of K are called *scalars*, and K is called the *field of scalars* of the vector space $(V,+,K,\cdot)$. The elements of V

are called *vectors*. The mapping $(\alpha, v) \to \alpha.v$ from $K \times V$ into V will be called *multiplication by scalars*.

From now on, the term "vector" will mean an element of a vector space. We will see vectors which do not resemble the vectors of physics in any way.

At the cost of some stylistic clumsiness in the formulation of many statements, we shall refer to the mapping $(\alpha, v) \to \alpha.v$ as multiplication by scalars, not as scalar multiplication. This is what the mapping really is. It is multiplication of vectors by scalars, not a multiplication whose results are scalars. We will usually omit $\cdot$ and write αv instead of $\alpha.v$.

Strictly speaking, a vector space is an ordered quadruple $(V, +, K, \cdot)$. However, as in the case of groups and rings, we shall usually refer to the *set* V as a vector space over K. If the field of scalars is fixed throughout a discussion, we shall speak of vector spaces, without reference to the field of scalars.

It will be convenient to think of a vector space as an abelian group with an additional structure on it supplied by the multiplication by scalars. The wording of Definition 39.1 was chosen to emphasize this point of view.

39.2 Examples: (a) Let $V = \mathbb{R} \oplus \mathbb{R}$ be the direct sum of two copies of $\mathbb{R}$ and let $K = \mathbb{R}$. We define multiplication by scalars in the most natural way: $\alpha(\beta,\gamma) = (\alpha\beta, \alpha\gamma)$ (for all $\alpha \in \mathbb{R}$, $(\beta,\gamma) \in V$).
It is easily seen that V is an $\mathbb{R}$-vector space.

(b) The same construction can be carried out with n-tuples of elements from any field K. Let K be a field and let $V = K \oplus K \oplus \ldots \oplus K$ be the direct sum of n copies of K, which is an abelian group under componentwise addition. We define multiplication by scalars also componentwise:
$\alpha(\beta_1, \beta_2, \ldots, \beta_n) = (\alpha\beta_1, \alpha\beta_2, \ldots, \alpha\beta_n)$ (for all $\alpha \in K$, $(\beta_1, \beta_2, \ldots, \beta_n) \in V$).
It is easily verified that V is a K-vector space. It will be designated by K^n, and will be called the *K-vector space of n-tuples*.

(c) Let V be the set of all real-valued functions defined on the interval $[0,1]$. For any two functions f,g in V, we define a new function $f+g$ in V by $(f+g)(x) = f(x) + g(x)$ for all $x \in [0,1]$.
V is an abelian group under this addition (called the *pointwise addition* of functions). Now let $K = \mathbb{R}$ and define a *pointwise multiplication* by scalars: $(\alpha f)(x) = \alpha f(x)$ for all $x \in [0,1]$, $\alpha \in \mathbb{R}$.
Then V is a vector space over $\mathbb{R}$ (cf. Example 29.2(i)).

When we put $(\alpha f)(x) = \alpha f(x)$ for all $x \in [0,1]$ $\alpha \in \mathbb{C}$, then V would not be a vector space over $\mathbb{C}$, because αf would not belong to V for all $\alpha \in \mathbb{C}, f \in V$, as the function αf is not real-valued when α is a complex number with a nonzero imaginary part.

(d) Let K be a field and let $K[x]$ be the ring of all polynomials over K. Let us forget that we can multiply two polynomials and concentrate on the fact that we can add them and multiply them by the elements of K (which are polynomials of degree zero, or the zero polynomial). It is easily seen that $K[x]$ is a K-vector space.

(e) Let n be a fixed natural number. Let K be a field and let V be the set of all polynomials over K which have degree n. Is V a vector space over K? No, because the sum of two polynomials of degree n is not always a polynomial of degree n (when the leading coefficients are opposites of each other). On the other hand, the set consisting of the zero polynomial and of all polynomials over K whose degrees are less than or equal to n is a vector space over K.

(f) Let V be a vector space over a field K, and let K_1 be a field contained in K (in this case, K_1 is called a *subfield of* K). Then V is a vector space over K_1, too, since the requirements in Definition 39.1 are satisfied by the elements of K_1 if they are satisfied by the elements of K.

(g) Let K be a field. When we define the multiplication by scalars as the multiplication in the field K, then K becomes a vector space over K. The conditions in Definition 39.1 are simply the distributivity laws, the associativity of multiplication and the very definition of the identity element in K.

(h) It follows from Example 39.2(f) and Example 39.2(g) that, if K_1 and K are fields such that $K_1 \subseteq K$, then K is a vector space over K_1: any field is a

vector space over its subfields. For instance, $\mathbb{C}$ is a vector space over $\mathbb{R}$, and $\mathbb{R}$ is a vector space over $\mathbb{Q}$.

39.3 Remarks: (1) A vector space is an abelian group and has an identity element, which we call zero and denote by 0. The underlying field K has a zero element, too, which is also denoted by 0. The reader should carefully distinguish between these zeroes. One of them is a vector, the other is a scalar. The vector zero is sometimes denoted by $\vec{0}$.

(2) Multiplication by scalars is a mapping from $K \times V$ into V, hence it is *not* a binary operation on V unless $K = V$. This feature distinguishes vector spaces from groups and rings. Multiplication and addition are binary operations on groups and rings.

Some basic facts are collected in the next lemma.

39.4 Lemma: *Let V be a vector space over a field K. For all $\alpha,\beta \in K$ and for all $u,v,w \in V$, the following hold.*

(1) $0 + v = v$.
(2) $-v + v = 0$.
(3) $-0 = 0$ (vector zero).
(4) $u + v = u + w$ *implies* $v = w$.
(5) $\alpha 0 = 0$.
(6) $0v = 0$.
(7) $\alpha(-v) = -(\alpha v) = (-\alpha)v$; *in particular,* $-1.v = -v$.
(8) $(\alpha - \beta)v = \alpha v - \beta v$.
(9) $\alpha(v - w) = \alpha v - \alpha w$.
(10) $\alpha v = 0$ *implies* $\alpha = 0$ or $v = 0$.
(11) $\alpha v = \beta v$ *implies* $\alpha = \beta$ or $v = 0$.
(12) $\alpha v = \alpha w$ *implies* $\alpha = 0$ or $v = w$.

Proof: (1),(2),(3),(4) hold in any group (Lemma 7.3, Lemma 8.2), also in the abelian group $(V,+)$.

(5) We are to prove $\alpha 0 = 0$ (vector zero). We observe

$$\alpha 0 + 0 = \alpha 0 = \alpha(0 + 0) = \alpha 0,$$

hence $\alpha 0 = 0$ by (4).

(6) We are to prove $0v = 0$ (on the left hand side, we have the scalar zero, on the right hand side, the vector zero). We observe

$$0v + 0 = 0v = (0 + 0)v = 0v + 0v,$$

hence $0v = 0$ by (4).

(7) We have $\quad 0 = \alpha 0 = \alpha(v + (-v)) = \alpha v + \alpha(-v)$

and $\quad 0 = 0v = (\alpha + (-\alpha))v = \alpha v + (-\alpha)v$

by (5) and (6), so $(-\alpha)v$ and $\alpha(-v)$ are the opposite of αv. Thus

$$\alpha(-v) = -(\alpha v) = (-\alpha)v.$$

(8) We are to show $(\alpha - \beta)v = \alpha v - \beta v$. Here $\alpha - \beta$ is an abbreviation for $\alpha + (-\beta)$ in K, and $\alpha v - \beta v$ is an abbreviation for $\alpha v + (-(\beta v))$ in V. We have indeed: $(\alpha - \beta)v = (\alpha + (-\beta))v = \alpha v + (-\beta)v = \alpha v + (-(\beta v)) = \alpha v - \beta v$.

(9) $\alpha(v - w) = \alpha(v + (-w)) = \alpha v + \alpha(-w) = \alpha v + (-(\alpha w)) = \alpha v - \alpha w$.

(10) Assume $\alpha v = 0$. If $\alpha \neq 0$, then α^{-1} exists in K and we get

$$v = 1v = (\alpha^{-1}\alpha)v = \alpha^{-1}(\alpha v) = \alpha^{-1}0 = 0.$$

(11) This follows from (8) and (10).

(12) This follows from (9) and (10). □

39.5 Lemma: *Let V be a vector space over a field K. Then, for all $\alpha_1, \alpha_2, \ldots, \alpha_n, \alpha$ in K and $v_1, v_2, \ldots, v_n, v$ in V, there hold*

$$(\alpha_1 + \alpha_2 + \cdots + \alpha_n)v = \alpha_1 v + \alpha_2 v + \cdots + \alpha_n v$$

and

$$\alpha(v_1 + v_2 + \cdots + v_n) = \alpha v_1 + \alpha v_2 + \cdots + \alpha v_n.$$

Proof: This follows by induction on n. The details are left to the reader. □

Just as there may be different group structures on a set, there may also be different vector space structures on a set. Here is an example.

39.6 Example: Let $V := \mathbb{C} \oplus \mathbb{C}$. We define a multiplication $\circ$ of the elements of V by complex numbers by declaring

$$c \circ (a,b) = (\bar{c}a, \bar{c}b) \qquad \text{for all } c \in \mathbb{C},\ (a,b) \in V.$$

This multiplication makes the abelian group V into a $\mathbb{C}$-vector space:

$$
\begin{aligned}
(1)\quad c \circ [(a,b) + (d,e)] &= c \circ (a + d, b + e) \\
&= (\bar{c}(a + d), \bar{c}(b + e)) \\
&= (\bar{c}a + \bar{c}d, \bar{c}b + \bar{c}e) \\
&= (\bar{c}a, \bar{c}b) + (\bar{c}d, \bar{c}e) \\
&= c \circ (a,b) + c \circ (d,e), \\
(2)\qquad (c + f) \circ (a,b) &= (\overline{(c + f)}a, \overline{(c + f)}b) \\
&= (\bar{c}a + \bar{f}a, \bar{c}b + \bar{f}b) \\
&= (\bar{c}a, \bar{c}b) + (\bar{f}a, \bar{f}b) \\
&= c \circ (a,b) + f \circ (a,b), \\
(3)\qquad (cf)(ab) &= (\overline{cf}a, \overline{cf}b) \\
&= (\bar{c}\bar{f}a, \bar{c}\bar{f}b) \\
&= c \circ (\bar{f}a, \bar{f}b) \\
&= c \circ (f \circ (a,b)), \\
(4)\qquad 1 \circ (a,b) &= (\bar{1}a, \bar{1}b) = (1a, 1b) = (a,b)
\end{aligned}
$$

for all $(a,b),(d,e) \in V$, $c,f \in \mathbb{C}$. Thus $(V,+,\mathbb{C},\circ)$ is a vector space, with the same set V, the same addition $+$ on V, the same underlying field $\mathbb{C}$ as the vector space $(V,+,\mathbb{C},\cdot)$ of Example 39.2(b) (in case $K = \mathbb{C}$, $n = 2$), but $(V,+,\mathbb{C},\circ)$ is distinct from $(V,+,\mathbb{C},\cdot)$ since the multiplication by scalars in these vector spaces are different.

Exercises

1. Determine whether $\mathbb{R} \times \mathbb{R}$ is an $\mathbb{R}$-vector space when
$(a,b) + (c,d) = (a + c, b + d)$, $a(c,d) = (c,d)$ for all $a,b,c,d \in \mathbb{R}$.

2. Determine whether $\mathbb{Q} \times \mathbb{Q}$ is a $\mathbb{Q}$-vector space when
$(a,b) + (c,d) = (a + c, 0)$, $a(c,d) = (ac,ad)$ for all $a,b,c,d \in \mathbb{Q}$.

3. Determine whether $\mathbb{Z}_7 \times \mathbb{Z}_7$ is a $\mathbb{Z}_7$-vector space when
$(a,b) + (c,d) = (a + c, b + d)$, $a(c,d) = (ac,0)$ for all $a,b,c,d \in \mathbb{Z}_7$.

4 Determine whether the set S of all sequences of real numbers is a vector space over $\mathbb{R}$ if addition and multiplication by scalars are defined by $(a_n) + (b_n) = (a_n + b_n), \quad a(b_n) = (ab_n)$
for all $(a_n),(b_n) \in S, a \in \mathbb{R}$ (here (a_n) is the sequence $a_1,a_2,a_3, \ldots$).

5. If $q \in \mathbb{N}$ and K is a field of q elements, how many elements does K^n have?

6. Construct a vector space with exactly four elements.

§40
Subspaces

Just as we defined subgroups and subrings, we will define sub(vector space)s. We contract this awkward expression into "subspace".

40.1 Definition: Let V be a vector space over a field K. A nonempty subset W of V is called a *subspace of* V if W itself is a K-vector space (under the addition and multiplication by scalars inherited from V).

A subspace W of V is an abelian group, thus a subgroup of $(V,+)$. Also, products by scalars of the element of W belong to W, so that $\alpha w \in W$ whenever $\alpha \in K$ and $w \in W$. Conversely, any subgroup W of V such that $\alpha w \in W$ for all $\alpha \in K$ and $w \in W$ is easily seen to be a subspace of V, for the conditions in Definition 39.1 are automatically satisfied for all elements of W if they are satisfied for all elements of V. Thus W is a subspace of V if and only if

(1) W is a subgroup of V under addition,

(ii) if $\alpha \in K$ and $w \in W$, then $\alpha w \in W$ (i.e.,W is closed under multiplication by scalars).

Here (1) embraces two conditions: (i) W is closed under addition, (ii') for any $w \in W$, the opposite $-w$ of w also belongs to W. Thus W is a subspace if and only if (i),(ii') and (ii) hold. One checks easily that (ii) implies (ii'): if (ii) holds and $w \in W$, then $(-1)w \in W$, hence $-w \in W$ by Lemma 39.4(7), so (ii') holds. Thus (ii') is superfluous. We proved the following lemma.

40.2 Lemma (Subspace criterion): *Let V be a vector space over a field K and let W be a nonempty subset of V. Then W is a subspace of V if and only if*

(i) $w_1 + w_2 \in W$ *for all* $w_1, w_2 \in W$,

(ii) $\alpha w \in W$ *for all* $\alpha \in K, w \in W$. □

So a nonempty subset of a vector space V is a subspace of V if and only if it is closed under addition and multiplication by scalars. The two closure properties of Lemma 40.2 can be combined to a single one. When (i) and (ii) of Lemma 40.2 hold, then

$$\alpha w_1 + \beta w_2 \in W \qquad \text{for all } \alpha,\beta \in K,\ w_1,w_2 \in W \qquad (*)$$

since $\alpha w_1, \beta w_2 \in W$ by (ii) and $\alpha w_1 + \beta w_2 \in W$ by (i). Conversely, if (*) holds, then, choosing $\alpha = 1$, $\beta = 1$, we see that (i) holds and, choosing $\beta = 0$, we see that (ii) holds. Thus (i) and (ii) are together equivalent to (*). Then we obtain another version of Lemma 40.2.

40.3 Lemma (Subspace criterion): *Let V be a vector space over a field K and let W be a nonempty subset of V. Then W is a subspace of V if and only if*

$$\alpha w_1 + \beta w_2 \in W \qquad \textit{for all } \alpha,\beta \in K,\ w_1,w_2 \in W. \qquad \square$$

The expression $\alpha w_1 + \beta w_2 \in W$ is said to be a linear combination of the vectors w_1,w_2. More generally, we have the

40.4 Definition: Let $v_1,v_2,\ldots,v_n$ be finitely many (not necessarily distinct) vectors of a vector space V over a field K. A vector of the form

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n,$$

where $\alpha_1,\alpha_2,\ldots,\alpha_n \in K$, is called a *$K$-linear combination* of the vectors $v_1,v_2,\ldots,v_n$. (If the underlying field K is clear from the context, we use the term "linear combination", without mentioning K.)

40.5 Lemma: *Let V be a vector space over a field K and let W be a subspace of V. If $w_1,w_2,\ldots,w_m$ are vectors in W, then every K-linear combination of these vectors belongs to W.*

Proof: This follows from Lemma 40.3 by induction on m. $\square$

40.6 Examples: (a) Let V be any vector space. Then $\{0\}$ and V are subspaces of V.

(b) Consider the vector space K^3 over a field K and put

$$W = \{(\lambda,\mu,\nu) \in K^3 : \nu = 0\} \subseteq K^3.$$

If $(\lambda_1,\mu_1,0)$ and $(\lambda_2,\mu_2,0)$ are arbitrary vectors in W and if α,β are arbitrary scalars, then $\alpha(\lambda_1,\mu_1,0) + \beta(\lambda_2,\mu_2,0) = (\alpha\lambda_1 + \beta\lambda_2, \alpha\mu_1 + \beta\mu_2, 0)$ belongs to W. By Lemma 40.3, W is a subspace of K^3.

(c) Consider the vector space K^3 over a field K and put

$$U = \{(\lambda,\mu,\nu) \in K^3 : \nu = 1\} \subseteq K^3.$$

Then $(0,0,1) \in U$, $(1,0,1) \in U$, but $(0,0,1) + (1,0,1) = (1,0,1+1) \notin U$ (why?) and U is not closed under addition. So U is not a subspace of K^3.

(d) Consider $\mathbb{R}^3$ over $\mathbb{R}$ and let

$$A = \{(\lambda,\mu,\nu) \in \mathbb{R}^3 : \nu \geqslant 0\} \subseteq \mathbb{R}^3.$$

If (λ_1,μ_1,ν_1) and (λ_2,μ_2,ν_2) are in A, then $\nu_1 \geqslant 0$, $\nu_2 \geqslant 0$, so $\nu_1 + \nu_2 \geqslant 0$ and

$$(\lambda_1,\mu_1,\nu_1) + (\lambda_2,\mu_2,\nu_2) = (\lambda_1 + \lambda_2, \mu_1 + \mu_2, \nu_1 + \nu_2)$$

belongs to A. Thus A is closed under addition. However, A is not a subspace of $\mathbb{R}^3$, since, for instance, $(0,0,1) \in A$ but $(-1)(0,0,1) \notin U$. This example shows that a subset of a vector space can be closed under addition without being closed under multiplication by scalars.

(e) Consider the vector space K^2 over a field K and let

$$M = \{(\lambda,\mu) \in K^2 : \lambda = 0 \text{ or } \mu = 0\} \subseteq K^2.$$

If $\alpha \in K$ and $(\lambda,\mu) \in M$, then $\lambda = 0$ or $\mu = 0$, so $\alpha\lambda = 0$ or $\alpha\mu = 0$, so $\alpha(\lambda,\mu) = (\alpha\lambda,\alpha\mu)$ belongs to M. Thus A is closed under multiplication by scalars. However, M is not a subspace of K^2, since, for instance, $(1,0)$, $(0,1) \in M$, but $(1,0) + (0,1) \notin M$. This example shows that a subset of a vector space can be closed under multiplication by scalars without being closed under addition.

(f) Consider the vector space K^2 over a field K and let γ,δ be two arbitrary but fixed elements of K. Put

$R = \{(\lambda,\mu) \in K^2 : \gamma\lambda + \delta\mu = 0\} \subseteq K^2$. Then R is a subspace of K^2:

(i) If (λ_1,μ_1), $(\lambda_2,\mu_2) \in R$, then $\gamma\lambda_1 + \delta\mu_1 = 0 = \gamma\lambda_2 + \delta\mu_2$, so $(\gamma\lambda_1 + \delta\mu_1) + (\gamma\lambda_2 + \delta\mu_2) = 0$, so $\gamma(\lambda_1 + \lambda_2) + \delta(\mu_1 + \mu_2) = 0$, so $(\lambda_1,\mu_1) + (\lambda_2,\mu_2) = (\lambda_1 + \lambda_2, \mu_1 + \mu_2) \in R$.

(ii) If $\alpha \in K$ and $(\lambda,\mu) \in R$, then $\gamma\lambda + \delta\mu = 0$, so $\gamma\alpha\lambda + \delta\alpha\mu = 0$, so $\alpha(\lambda,\mu) = (\alpha\lambda,\alpha\mu) \in R$.

(g) Let V be a vector space over a field K and let $\{W_i: i \in I\}$ be a collection of subspaces of V. Then their intersection $W := \bigcap_{i \in I} W_i$ is a subspace of V. First of all, this intersection is not empty, since $0 \in W_i$ for all $i \in I$. Also, if $\alpha,\beta \in K$ and $w_1,w_2 \in W$, then $\alpha,\beta \in K$ and $w_1,w_2 \in W_i$ for all $i \in I$, so $\alpha w_1 + \beta w_2 \in W_i$ for all $i \in I$, so $\alpha w_1 + \beta w_2 \in W$.

(h) Let V be the $\mathbb{R}$-vector space of all real-valued functions defined on $[0,1]$ (See Example 39.2(c)) and let α be a fixed number in $[0,1]$. We put

$$T_\alpha = \{f \in V: f \text{ is continuous at } \alpha\}.$$

It is known from analysis that, if f and g are functions, continuous at α, then $f + g$ is also continuous at α. If f is continuous at α and $\beta \in \mathbb{R}$, then βf is continuous at α. Hence T_α is a subspace of V.

(i) Let V be the $\mathbb{R}$-vector space of all real-valued functions defined on $[0,1]$. We put

$$C([0,1]) = \{f \in V: f \text{ is continuous on } [0,1]\}.$$

We know from analysis that, if f and g are functions, continuous on $[0,1]$, then $f + g$ is also continuous on $[0,1]$. If f is continuous on $[0,1]$ and $\beta \in \mathbb{R}$, then βf is continuous on $[0,1]$. Hence $C([0,1])$ is a subspace of V. This conclusion can be drawn also by observing that $C([0,1]) = \bigcap_{\alpha \in [0,1]} T_\alpha$ and appealing to Example 40.6(g) and Example 40.6(h).

(j) Let $C^1([0,1]) = \{f \in C([0,1]): f' \text{ exists and is continuous on } [0,1]\}$. $C^1([0,1])$ is a nonempty subset of $C([0,1])$. If $\alpha,\beta \in \mathbb{R}$ and $f,g \in C^1([0,1])$, then, as is well known from analysis, $(\alpha f + \beta g)'$ exists, is equal to $\alpha f' + \beta g'$ and is continuous on $[0,1]$. Hence $\alpha f + \beta g \in C^1([0,1])$ and therefore $C^1([0,1])$ is a subspace of $C([0,1])$.

Similarly, for $k \in \mathbb{N}$, we put

$$C^k([0,1]) = \{f \in C([0,1]): f^{(k)} \text{ exists and is continuous on } [0,1]\}.$$

Since the existence of the k-th derivative of f implies the existence and continuity of the first $k - 1$ derivatives $f', f'', \ldots, f^{(k-1)}$, we see $C^k([0,1])$ is a subset of $C^{k-1}([0,1])$. From the formula

$$(\alpha f + \beta g)^{(k)} = \alpha f^{(k)} + \beta g^{(k)} \qquad (\alpha,\beta \in \mathbb{R}, f,g \in C^k([0,1]))$$

it is easily seen that $C^k([0,1])$ is a subspace of $C([0,1])$ and of $C^{k-1}([0,1])$.

We write $C^\infty([0,1]) = \bigcap_{k \in \mathbb{N}} C^k([0,1])$. From Example 40.6(g), we infer that $C^\infty([0,1])$ is also a subspace of $C([0,1])$ and of each $C^k([0,1])$.

(k) Let $p(x)$ and $q(x)$ be continuous functions, defined on $[0,1]$. We write

$$L = \{f \in C^2([0,1]): f''(x) + p(x)f'(x) + q(x)f(x) = 0 \text{ for all } x \in [0,1]\}.$$

L is a nonempty subset of $C^2([0,1])$. If $\alpha,\beta \in \mathbb{R}$ and $f,g \in L$, then

$$\begin{aligned}&(\alpha f + \beta g)''(x) + p(x)(\alpha f + \beta g)'(x) + q(x)(\alpha f + \beta g)(x)\\ &= \alpha f''(x) + \beta g''(x) + p(x)\alpha f'(x) + p(x)\beta g'(x) + q(x)\alpha f(x) + q(x)\beta g(x)\\ &= \alpha(f''(x) + p(x)f'(x) + q(x)f(x)) + \beta(g''(x) + p(x)g'(x) + q(x)g(x))\\ &= \alpha 0 + \beta 0 = 0\end{aligned}$$

for all $x \in [0,1]$, so $\alpha f + \beta g \in L$. Thus L is a subspace of $C^2([0,1])$.

(l) So far, we spoke of subspaces without referring to the underlying field. Sometimes it might be necessary to mention the underlying field. Let $V = \mathbb{C}^2$ be the $\mathbb{C}$-vector space of ordered pairs of complex numbers. Then V is an $\mathbb{R}$-vector space, too (Example 39.2(f)). We put

$$W = \{(\lambda,\overline{\lambda}): \lambda \in \mathbb{C}\} = \{(\lambda,\mu): \lambda \in \mathbb{C}, \mu = \overline{\lambda}\},$$

where $\overline{\ }$ denotes complex conjugation. If (λ,μ), $(\nu,\rho) \in W$ and $\alpha,\beta \in \mathbb{R}$, then $\mu = \overline{\lambda}$ and $\rho = \overline{\nu}$, so $\alpha(\lambda,\mu) + \beta(\nu,\rho) = (\alpha\lambda + \beta\nu, \alpha\mu + \beta\rho)$ with

$$\begin{aligned}\overline{\alpha\lambda + \beta\nu} &= \overline{\alpha}\overline{\lambda} + \overline{\beta}\overline{\nu}\\ &= \overline{\alpha}\,\overline{\lambda} + \overline{\beta}\,\overline{\nu}\\ &= \alpha\overline{\lambda} + \beta\overline{\nu} \qquad (c)\\ &= \alpha\mu + \beta\rho,\end{aligned}$$

and $\alpha(\lambda,\mu) + \beta(\nu,\rho) \in W$. Thus W is a subspace of the $\mathbb{R}$-vector space V. However, W is not a subspace of the $\mathbb{C}$-vector space V, for the critical equation (c) need not be true when α,β are complex numbers (with nonzero imaginary parts). We may say W is an $\mathbb{R}$-subspace of V, but not a $\mathbb{C}$-subspace of V.

40.7 Theorem: *Let V be a vector space over a field K and let $A = \{v_1,v_2,\ldots,v_n\}$ be a finite nonempty subset of V. Then the set*

$$W = \{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n\}$$

of all linear combinations of the vectors $v_1,v_2,\ldots,v_n$ is a subspace of V.

Proof: Since A is not empty, $W \neq \emptyset$. If $\alpha,\beta \in K$ and $u,w \in W$, then

$$u = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n, \quad w = \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_n v_n$$

with suitable $\alpha_1,\alpha_2,\ldots,\alpha_n, \beta_1,\beta_2,\ldots,\beta_n \in K$ and

$$\begin{aligned}\alpha u + \beta w &= \alpha(\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n) + \beta(\beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_n v_n)\\ &= (\alpha\alpha_1 + \beta\beta_1)v_1 + (\alpha\alpha_2 + \beta\beta_2)v_2 + \cdots + (\alpha\alpha_n + \beta\beta_n)v_n\end{aligned}$$

belongs to W. Hence W is a subspace of V (Lemma 40.3). [Notice that $v_1, v_2, \ldots, v_n$ are not assumed to be distinct.] □

We extend this theorem to infinite subsets of V.

40.8 Theorem: *Let V be a vector space over a field K and let $A = \{v_i : i \in I\}$ be a (finite or infinite) nonempty subset of V. Then the set*

$$W = \{\alpha_1 v_{i_1} + \alpha_2 v_{i_2} + \cdots + \alpha_n v_{i_n} \in V : \alpha_1, \alpha_2, \ldots, \alpha_n \in K,\ v_{i_1}, v_{i_2}, \ldots, v_{i_n} \in A,\ n \in \mathbb{N}\}$$

of all finite linear combinations of the vectors in A is a subspace of V.

Proof: Since A is not empty, $W \neq \varnothing$. If $\alpha, \beta \in K$ and $u, w \in W$, then

$$u = \alpha_1 v_{i_1} + \alpha_2 v_{i_2} + \cdots + \alpha_n v_{i_n}, \quad w = \beta_1 v_{j_1} + \beta_2 v_{j_2} + \cdots + \beta_m v_{j_m}$$

with suitable $\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_n \in K$, $v_{i_1}, v_{i_2}, \ldots, v_{i_n}, v_{j_1}, v_{j_2}, \ldots, v_{j_m} \in A$, $n, m \in \mathbb{N}$ and

$$\begin{aligned} \alpha u + \beta w &= \alpha(\alpha_1 v_{i_1} + \alpha_2 v_{i_2} + \cdots + \alpha_n v_{i_n}) + \beta(\beta_1 v_{j_1} + \beta_2 v_{j_2} + \cdots + \beta_m v_{j_m}) \\ &= \alpha\alpha_1 v_{i_1} + \alpha\alpha_2 v_{i_2} + \cdots + \alpha\alpha_n v_{i_n} + \beta\beta_1 v_{j_1} + \beta\beta_2 v_{j_2} + \cdots + \beta\beta_m v_{j_m} \end{aligned}$$

is a K-linear combination of the vectors $v_{i_1}, v_{i_2}, \ldots, v_{i_n}, v_{j_1}, v_{j_2}, \ldots, v_{j_m}$ in A. So $\alpha u + \beta w$ belongs to W and W is a subspace of V. □

40.9 Definition: The subspace W of Theorem 40.7 or Theorem 40.8 is called the *K-span of A*, or the *K-span of the vectors in A*, or the *subspace spanned by the (vectors in) A*. It will be denoted by $s_K(A)$. In case $A = \{v_1, v_2, \ldots, v_n\}$ is a finite set, we write $s_K(v_1, v_2, \ldots, v_n)$ instead of $s_K(\{v_1, v_2, \ldots, v_n\})$. By convention, we put $s_K(\varnothing) = \{0\}$. When there is no need to refer to the field K of scalars, we speak of the *span* of A, and denote it by $s(A)$.

The next lemma justifies the convention $s_K(\varnothing) = \{0\}$.

40.10 Lemma: *Let V be a vector space over a field K and let $A \subseteq V$. Then $s_K(A)$ is the smallest subspace of V which contains A. More exactly, if U is a subspace of V and $A \subseteq U$, then $s_K(A) \subseteq U$.*

Proof: If $A = \varnothing$, then $s_K(A) = \{0\} \subseteq U$ for any subspace U of V and the theorem is proved in this case. Suppose now $A \neq \varnothing$. If $A \subseteq U$ and U is a subspace of V, then every linear combination of the vectors in A belongs to U by Lemma 40.5. Hence $s(A) \subseteq U$. □

40.11 Lemma: *Let V be a vector space over a field K and let A,B be subspaces of V such that $A \subseteq s(B)$ and $B \subseteq s(A)$. Then $s(A) = s(B)$.*

Proof: Since $A \subseteq s(B)$ and $s(B)$ is a subspace of V, we have $s(A) \subseteq s(B)$ by Lemma 40.10. In like manner, since $B \subseteq s(A)$ and $s(A)$ is a subspace of V, we get $s(B) \subseteq s(A)$. Thus $s(A) = s(B)$. □

40.12 Examples: **(a)** Let V be a vector space over a field K and let A be a subset of V having only one element, say $A = \{v\}$. Then the span $s(v)$ of A is the set

$$\{\alpha v \in V: \alpha \in K\}$$

of all scalar multiples of v. In case $K = \mathbb{R}$ and $V = \mathbb{R}^2$ or $V = \mathbb{R}^3$, this span is usually identified with the line through the origin determined by v.

(b) Let V be a vector space over a field K and let u,v be two vectors in V. The span $s(u,v)$ of these vectors is

$$\{\alpha u + \beta v \in V: \alpha,\beta \in K\}.$$

In case v is a scalar multiple γu of u, we have

$$s(u,v) = \{\alpha u + \beta v \in V: \alpha,\beta \in K\} = \{(\alpha + \beta\gamma)u: \alpha,\beta \in K\} = \{\delta u: \delta \in K\} = s(u).$$

We see it is possible that $A \subset B$ and $s(A) = s(B)$. In case $K = \mathbb{R}$ and $V = \mathbb{R}^3$ and v is not a scalar multiple of u, this span is usually identified with the plane through the origin determined by u and v.

(c) In the vector space $\mathbb{R}^2$ over $\mathbb{R}$, consider the set

$$A = \{(1,0),(2,0),\ldots,(10000,0)\}.$$

The span $s(A)$ is easily seen to be $\{(a,0) \in \mathbb{R}^2 : a \in \mathbb{R}\}$, which is also the span of $\{(1,0)\} \subseteq \mathbb{R}^2$. Thus the number of vectors in A may be large, but this does not imply that $s(A)$ is a "big" subspace.

(d) Let V be a vector space over a field K and let A,B be subsets of V with $A \subseteq B$. Then $A \subseteq B \subseteq s(B)$ and, since $s(B)$ is a subspace of V, Lemma 40.10 yields $s(A) \subseteq s(B)$. So $A \subseteq B$ implies $s(A) \subseteq s(B)$. We have seen in Example 40.12(b) and Example 40.12(c) that $A \subset B$ does not necessarily imply $s(A) \subset s(B)$.

Exercises

1. Let V be a vector space over a field K. If W is a subspace of V and U is a subspace of W, prove that U is a subspace of V.

2. Prove that the set of all sequences of real numbers converging to 0 is a subspace of the $\mathbb{R}$-vector space S (see § 39, Ex. 5). What do you say about the set of all convergent sequences, all bounded sequences, all monotonic sequences, and all sequences with at most finitely many nonzero terms?

3. Consider the $\mathbb{R}$-vector space V of Example 39.2(c). Determine whether the following are subspaces of V: the set of bounded functions, the set of even functions, the set of integrable functions, the set of monotonic functions, the set of functions with at most finitely many points of discontinuity (all with domains $[0,1]$).

4. Determine whether
$$\{(\alpha,\beta,\gamma) \in \mathbb{R}^3 : 5\alpha - 4\beta + 2\gamma = 0\}$$
$$\{(\alpha,\beta,\gamma) \in \mathbb{R}^3 : 5\alpha - 4\beta + 2\gamma \geqslant 0\}$$
$$\{(\alpha,\beta,\gamma) \in \mathbb{Z}_{11}{}^3 : 5\alpha - 4\beta + 2\gamma = 0\}$$
are subspaces of the vector spaces indicated.

5. Is $(1,0,1) \in \mathbb{R}^3$ in the $\mathbb{R}$-span of $\{(5,4,1), (3,2,2)\} \subseteq \mathbb{R}^3$?

§41
Factor Spaces

In the preceding paragraph, we discussed subspaces, which are the analogues of subgroups and subrings. We now wish to discuss the analogues of factor groups and factor rings.

Let V be a vector space over a field K and let W be a subspace of V. Then W is a subgroup of the additive group V, and we can build the factor group V/W. The elements of V/W are cosets $v + W$, where $v \in V$; the sum of two cosets $v_1 + W$ and $v_2 + W$ is the coset $(v_1 + v_2) + W$. The operation on V/W is denoted by "+", but "+" designates in V/W an operation distinct from the addition in V. The question arises: is it possible to define on V/W a kind of multiplication by scalars so that V/W becomes a vector space over K? The most natural multiplication $\circ$ is to put

$$\alpha \circ (v + W) = \alpha v + W \qquad \text{for all } \alpha \in K,\ v + W \in V/W.$$

We prove that $\circ$ is well defined. To this end, we must show that the implication

$$u + W = v + W \implies \alpha u + W = \alpha v + W \qquad (\text{for all } \alpha \in K,\ v, u \in V)$$

is valid. This implication is equivalent to

$$u - v \in W \implies \alpha u + W = \alpha v + W$$

hence to

$$u - v \in W \implies \alpha u - \alpha v \in W.$$

Since W is a subspace of V, it is closed under multiplication by scalars, hence $\alpha(u - v) \in W$ whenever $u - v \in W$. This proves that the above multiplication $\circ$ by scalars is well defined.

It is now quite straightforward to show that $(V/W, +, K, \circ)$ is a vector space. For any $\alpha, \beta \in K$, $u, v \in V$, we have

$$(1) \qquad \begin{aligned} \alpha \circ ((u + W) + (v + W)) &= \alpha \circ ((u + v) + W) \\ &= \alpha(u + v) + W \end{aligned}$$

$$= (\alpha u + \alpha v) + W$$
$$= (\alpha u + W) + (\alpha v + W)$$
$$= \alpha \circ (u + W) + \alpha \circ (v + W),$$

(2) $$(\alpha + \beta) \circ (u + W) = (\alpha + \beta)u + W$$
$$= (\alpha u + \beta u) + W$$
$$= (\alpha u + W) + (\beta u + W)$$
$$= \alpha \circ (u + W) + \beta \circ (u + W),$$

(3) $$(\alpha\beta) \circ (u + W) = (\alpha\beta)u + W$$
$$= \alpha(\beta u) + W$$
$$= \alpha \circ (\beta u + W)$$
$$= \alpha \circ (\beta \circ (u + W)),$$

(4) $$1 \circ (u + W) = 1u + W$$
$$= u + W.$$

Thus $(V/W, +, K, \circ)$ is a vector space.

We employed the symbol "$\circ$" chiefly to emphasize that multiplication of the elements in V/W by scalars is distinct from the multiplication of the elements in V by scalars. For ease of notation, we shall drop "$\circ$" and write simply $\alpha(u + W)$ instead of $\alpha \circ (u + W)$. Also, we will write V/W for $(V/W, +, K, \circ)$. The following theorem summarizes this discussion.

41.1 Theorem: *Let V be a vector space over a field K and let W be a subspace of V. Then the abelian group V/W is a vector space over K if multiplication by scalars is defined by*

$$\alpha(u + W) = \alpha u + W \qquad \text{for all } \alpha \in K, u \in V. \qquad \square$$

41.2 Definition: Let V be a vector space over a field K and let W be a subspace of V. The K-vector space V/W of Theorem 41.1 is called the *factor space of V by W*, or the *factor space V mod(ulo) W*.

We know that factor groups (rings) are closely related to homomorphisms of groups (rings). The same is true for factor spaces.

41.3 Definition: Let V and U be vector spaces over the same field K. A mapping $\varphi: V \to U$ is called a *vector space homomorphism*, or a *K-linear transformation*, or a *K-linear mapping* if

$$(v_1 + v_2)\varphi = v_1\varphi + v_2\varphi \quad \text{and} \quad (\alpha v)\varphi = \alpha(v\varphi)$$

for all $v_1, v_2, v \in V$, $\alpha \in K$. When there is no need to emphasize the field of scalars, we speak simply of linear transformations or linear mappings.

More exactly, when $(V,+,K,\cdot)$ and $(U,\oplus,K,\circ)$ are vector spaces, the mapping $\varphi: V \to U$ is a vector space homomorphism provided

$$(v_1 + v_2)\varphi = v_1\varphi \oplus v_2\varphi \quad \text{and} \quad (\alpha v)\varphi = \alpha \circ (v\varphi)$$

for all $v_1, v_2, v \in V$, $\alpha \in K$. Notice that the field of scalars of both vector spaces are the same. A linear transformation from V into U cannot be defined if V and U are vector spaces over different fields.

A mapping $\varphi: V \to U$ such that $(v_1 + v_2)\varphi = v_1\varphi + v_2\varphi$ for all $v_1, v_2 \in V$ is said to be *additive*. So an additive mapping is just a group homomorphism from the group $(V,+)$ into $(U,+)$. A mapping $\varphi: V \to U$ such that $(\alpha v)\varphi = \alpha(v\varphi)$ for all $v \in V$, $\alpha \in K$ is said to be *homogeneous*. A homogeneous mapping is one that preserves the multiplication by scalars. A mapping may be additive without being homogeneous, and it may be homogeneous without being additive. In order to be a linear transformation, a mapping shoud be both additive and homogeneous.

A vector space homomorphism is therefore a homomorphism of additive groups which preserves multiplication by scalars as well. This observation enables us to use the properties of group homomorphisms whenever we investigate vector space homomorphisms.

41.4 Lemma: *Let V and U be a vector spacer over a field K. A function $\varphi: V \to U$ is a K-linear mapping if and only if*

$$(\alpha v_1 + \beta v_2)\varphi = \alpha(v_1\varphi) + \beta(v_2\varphi)$$

for all $\alpha, \beta \in K$, $v_1, v_2 \in V$.

Proof: If φ is a K-linear mapping and $\alpha,\beta \in K$, $v_1,v_2 \in V$, then

$$(\alpha v_1 + \beta v_2)\varphi = (\alpha v_1)\varphi + (\beta v_2)\varphi = \alpha(v_1\varphi) + \beta(v_2\varphi)$$

since φ is additive and homogeneous. Conversely, if we have $(\alpha v_1 + \beta v_2)\varphi = \alpha(v_1\varphi) + \beta(v_2\varphi)$ for all $\alpha,\beta \in K$, $v_1,v_2 \in V$, then, choosing $\alpha = \beta = 1$, we see that φ is additive and choosing $\beta = 0$, we see that φ is homogeneous. □

41.5 Lemma: *Let V,U be a vector spacer over a field K and let $\varphi: V \to U$ be a vector space homomorphism.*

(1) $0\varphi = 0$.

(2) $(-v)\varphi = -(v\varphi)$ *for all* $v \in V$.

(3) $(\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n)\varphi = \alpha_1(v_1\varphi) + \alpha_2(v_2\varphi) + \cdots + \alpha_n(v_n\varphi)$ *for all* $\alpha_1,\alpha_2,\ldots,\alpha_n \in K$ *and for all* $v_1,v_2,\ldots,v_n \in V$.

(4) $(nv)\varphi = n(v\varphi)$ *for all* $n \in \mathbb{Z}$.

Proof: (1),(2),(4) follow respectively from (1),(2),(4) of Lemma 20.3 and (3) follows from Lemma 20.3(3) by the homogeneity of φ, or from Lemma 41.4 by induction on n. □

41.6 Examples: (a) Let K be a field and let $\varphi: K^3 \longrightarrow K^2$. Then

$$(\lambda,\mu,\nu) \to (\lambda,\mu)$$

$$\begin{aligned}(\alpha(\lambda,\mu,\nu) + \beta(\lambda',\mu',\nu'))\varphi &= ((\alpha\lambda,\alpha\mu,\alpha\nu) + (\beta\lambda',\beta\mu',\beta\nu'))\varphi \\ &= ((\alpha\lambda + \beta\lambda', \alpha\mu + \beta\mu', \alpha\nu + \beta\nu'))\varphi \\ &= (\alpha\lambda + \beta\lambda', \alpha\mu + \beta\mu') \\ &= (\alpha\lambda,\alpha\mu) + (\beta\lambda',\beta\mu') \\ &= \alpha(\lambda,\mu) + \beta(\lambda',\mu') \\ &= \alpha((\lambda,\mu,\nu))\varphi + \beta((\lambda',\mu',\nu'))\varphi\end{aligned}$$

for all $\alpha,\beta \in K$, (λ,μ,ν), $(\lambda',\mu',\nu') \in K^3$. Hence φ is a K-linear transformation.

(b) Let K be a field and let $\varphi: K^2 \longrightarrow K^2$. Then

$$(\lambda,\mu) \to (\mu,\lambda)$$

$$\begin{aligned}(\alpha(\lambda,\mu) + \beta(\lambda',\mu'))\varphi &= ((\alpha\lambda,\alpha\mu) + (\beta\lambda',\beta\mu'))\varphi = ((\alpha\lambda + \beta\lambda', \alpha\mu + \beta\mu'))\varphi \\ &= (\alpha\mu + \beta\mu', \alpha\lambda + \beta\lambda') = (\alpha\mu,\alpha\lambda) + (\beta\mu',\beta\lambda') = \alpha(\mu,\lambda) + \beta(\mu',\lambda') \\ &= \alpha((\lambda,\mu))\varphi + \beta((\lambda',\mu'))\varphi\end{aligned}$$

for all $\alpha,\beta \in K$, $(\lambda,\mu),(\lambda',\mu') \in K^2$. Hence φ is a vector space homomorphism.

(c) The mapping $\varphi: C^1([0,1]) \to \mathbb{R}$ is $\mathbb{R}$-linear, because

$$f \to f(\tfrac{1}{2})$$

$(\alpha f + \beta g)\varphi = (\alpha f + \beta g)(\frac{1}{2}) = (\alpha f)(\frac{1}{2}) + (\beta g)(\frac{1}{2}) = \alpha(f(\frac{1}{2})) + \beta(g(\frac{1}{2})) = \alpha(f\varphi) + \beta(g\varphi)$

for all $\alpha,\beta \in \mathbb{R}, f,g \in C^1([0,1])$. Likewise, for any $\gamma \in [0,1]$, the mapping

$$\varphi_\gamma: C^1([0,1]) \to \mathbb{R}$$
$$f \to f(\gamma)$$

is a vector space homomorphism.

(d) Let V,U be vector spaces over a field K and let W be a subspace of V. If $\varphi: V \to U$ is a vector space homomorphism, then its restriction

$$\varphi_W: W \to U$$

to W is also a vector space homomorphism, because

$$(\alpha w_1 + \beta w_2)\varphi = \alpha(w_1\varphi) + \beta(w_2\varphi)$$

for all $\alpha,\beta \in K$, $w_1,w_2 \in W$, as this holds in fact for all $\alpha,\beta \in K$, $w_1,w_2 \in V$ (Lemma 41.4).

(e) Let V,U be vector spaces over a field K and let K_1 be a field contained in K. Then V,U are vector spaces over K_1, too (Example 39.2(f)). If $\varphi: V \to U$ is a K-linear mapping, then φ is also a K_1-linear mapping, because

$$(\alpha w_1 + \beta w_2)\varphi = \alpha(w_1\varphi) + \beta(w_2\varphi)$$

for all $\alpha,\beta \in K_1$, $v_1,v_2 \in V$, as this holds in fact for all $\alpha,\beta \in K$, $v_1,v_2 \in V$ (Lemma 41.4).

(f) The mapping $T: C^2([0,1]) \to C([0,1])$ is a vector space homomorphism

$$y \to y'' - 5y' + 6y$$

because

$$\begin{aligned}(\alpha y_1 + \beta y_2)T &= (\alpha y_1 + \beta y_2)'' - 5(\alpha y_1 + \beta y_2)' + 6(\alpha y_1 + \beta y_2) \\ &= \alpha y_1'' + \beta y_2'' - 5(\alpha y_1' + \beta y_2') + 6(\alpha y_1 + \beta y_2) \\ &= \alpha(y_1'' - 5y_1' + 6y_1) + \beta(y_2'' - 5y_2' + 6y_2) \\ &= \alpha(y_1T) + \beta(y_2T)\end{aligned}$$

for any $\alpha,\beta \in \mathbb{R}$, $y_1,y_2 \in C^2([0,1])$. In the theory of ordinary differential equations, this mapping is called a *linear differential operator* and is usually denoted by $D^2 - 5D + 6$.

In the rest of this paragraph, we establish the counterparts of certain theorems discussed in §§ 20, 21.

41.7 Theorem: *Let V,U,W be a vector spaces over a field K. Let $\varphi: V \to U$ and $\psi: U \to W$ be vector space homomorphisms. Then the composition mapping $\varphi\psi: V \to W$ is a vector space homomorphism from V into W.*

Proof: $\varphi\psi$ is a group homomorphism (is additive) by Theorem 20.4. Also

$$(\alpha v)\varphi\psi = ((\alpha v)\varphi)\psi = (\alpha(v\varphi))\psi = \alpha((v\varphi)\psi) = \alpha(v(\varphi\psi))$$

for all $\alpha \in K$, $v \in V$, hence $\varphi\psi$ is homogeneous. Thus $\varphi\psi$ is a vector space homomorphism. □

41.8 Theorem: *Let V,U be a vector spaces over a field K and let $\varphi: V \to U$ be K-linear. Then $Im\,\varphi = \{v\varphi \in U: v \in V\}$ is a subspace of U and $Ker\,\varphi = \{v \in V: v\varphi = 0\}$ is a subspace of V.*

Proof: $Im\,\varphi$ is a subgroup of $(U,+)$ by Theorem 20.6. Also, if $u \in Im\,\varphi$ and $\alpha \in K$, then $u = v\varphi$ for some $v \in V$, so $\alpha u = \alpha(v\varphi) = (\alpha v)\varphi$, so $\alpha u \in Im\,\varphi$. Thus $Im\,\varphi$ is closed under multiplication by scalars. Therefore $Im\,\varphi$ is a subspace of U.

$Ker\,\varphi$ is a subgroup of $(V,+)$ by Theorem 20.6. Also, if $v \in Ker\,\varphi$ and $\alpha \in K$, then $v\varphi = 0$, so $(\alpha v)\varphi = \alpha(v\varphi) = \alpha 0 = 0$ by Lemma 39.4(6), so $\alpha v \in Ker\,\varphi$. Thus $Ker\,\varphi$ is closed under multiplication by scalars. Therefore $Ker\,\varphi$ is a subspace of V. □

41.9 Definition: Let V,U be a vector spaces over a field K. A vector space homomorphism $\varphi: V \to U$ is called a *vector space isomorphism* if φ is one-to-one and onto. If there is a vector space isomorphism from V onto U, we say V *is isomorphic to* U, and write $V \cong U$.

So a vector space isomorphism is an additive group isomorphism which preserves multiplication by scalars. We use the same symbol "$\cong$" for isomorphic vector spaces as for isomorphic groups. This will not lead to confusion. When there is any danger of confusion, we will state explicitly whether we mean vector space isomorphism or group isomorphism.

41.10 Lemma: *Let V,U,W be a vector spaces over a field K and let $\varphi: V \to U$ and $\varphi: U \to W$ be vector space isomorphisms.*
(1) *The composition $\varphi\psi: V \to W$ is a vector space isomorphism from V onto W.*
(2) *The inverse $\varphi^{-1}: U \to V$ of φ is a vector space isomorphism from U onto V.*

Proof: (1) $\varphi\psi$ is a vector space homomorphism by Theorem 41.7, and $\varphi\psi$ is one-to-one and onto by Theorem 3.13. So $\varphi\psi$ is a vector space isomorphism.

(2) $\varphi^{-1}: U \to V$ is an isomorphism of additive groups by Lemma 20.11(2). We have only to show that φ^{-1} preserves multiplication by scalars. Let $u \in U$ and $\alpha \in K$. Then $u = v\varphi$ for some uniquely determined $v \in V$, namely $v = u\varphi^{-1}$. Since $(\alpha v)\varphi = \alpha(v\varphi) = \alpha u$, we have $(\alpha u)\varphi^{-1} = \alpha v$. Thus $(\alpha u)\varphi^{-1} = \alpha v = \alpha(u\varphi^{-1})$. So φ^{-1} preserves multiplication by scalars. □

As in the case of groups, we see that

$$V \cong V,$$
$$\text{if } V \cong U, \text{ then } U \cong V,$$
$$\text{if } V \cong U \text{ and } U \cong W, \text{ then } V \cong W$$

for all K-vector spaces V,U,W, where K is any field. Thus $\cong$ is an equivalence relation, but me must refrain from saying "on the *set* of K-vector spaces".

In view of the symmetry property of $\cong$, it is legitimate to say that V and U are isomorphic when V is isomorphic to U.

41.11 Theorem: *Let V be a vector space over a field K and let W be a subspace of V. Then the mapping*

$$\nu: V \to V/W$$
$$v \to v + W$$

is a vector space homomorphism. It is onto V/W. Also, $Ker\ \nu = W$. (This mapping ν is called the *natural* or *canonical homomorphism from V onto V/W*).

Proof: ν is an additive group homomorphism from V onto V/W such that $Ker\ \nu = W$ (Theorem 20.12). Since $(\alpha v)\nu = \alpha v + W = \alpha(v + W) = \alpha(v\nu)$ for all $\alpha \in K$, $v \in V$, we see that ν is a vector space homomorphism. □

41.12 Theorem (Fundamental theorem on homomorphisms): *Let K be a field. Let V,V_1 be K-vector spaces and let $\varphi: V \to V_1$ be a vector space homomorphism. Let $W = Ker\ \varphi$ and let $\nu: V \to V/W$ be the associated natural homomorphism.*
Then there is a vector space homomorphism $\psi: V/W \to V_1$ such that

$$\nu\psi = \varphi.$$

Proof: From Theorem 20.15, we know that $\psi: V/W \to V_1$ is a well defined,

$$v + W \to v\varphi$$

one-to-one homomorphism of additive groups with $\nu\psi = \varphi$. For all $\alpha \in K$, $v \in V$, we have $(\alpha(v + W))\psi = (\alpha v + W)\psi = (\alpha v)\varphi = \alpha(v\varphi) = \alpha((v + W)\psi)$, so ψ is homogeneous and is therefore a vector space homomorphism. □

41.13 Theorem: *Let V,U be vector spaces over a field K and let $\varphi: V \to U$ be a vector space homomorphism. Then*

$$V/Ker\ \varphi \cong Im\ \varphi \quad (as\ vector\ spaces).$$

Proof: From Theorem 20.16 and its proof, we know that

$$\psi: V/Ker\ \varphi \to Im\ \varphi$$
$$v + Ker\ \varphi \to v\varphi$$

is an isomorphism of additive groups, thus $V/Ker\ \varphi \cong Im\ \varphi$ as groups; and ψ is a vector space homomorphism by Theorem 41.12. Hence $\psi: V/Ker\ \varphi \to Im\ \varphi$ is a vector space isomorphism and $V/Ker\ \varphi \cong Im\ \varphi$ as vector spaces. □

41.14 Theorem: *Let V,V_1 be vector spaces over a field K and let $\varphi: V \to V_1$ be a vector space homomorphism from V onto V_1.*
(1) *Each subspace W of V with $Ker\ \varphi \subseteq W$ is mapped to a subspace of V_1, which will be denoted by W_1.*
(2) *If W,U are subspaces of V with $Ker\ \varphi \subseteq W \subseteq U$, then $W_1 \subseteq U_1$.*

(3) *If W,U are subspaces of V with $Ker\ \varphi \subseteq W$ and $Ker\ \varphi \subseteq U$, and if $W_1 \subseteq U_1$, then $W \subseteq U$.*

(4) *If W,U are subspaces of V with $Ker\ \varphi \subseteq W$ and $Ker\ \varphi \subseteq U$, and if $W_1 = U_1$, then $W = U$.*

(5) *If S is any subspace of V_1, then there is a subspace W of V such that $Ker\ \varphi \subseteq W$ and $W_1 = S$.*

(6) *If U is a subspace of V with $Ker\ \varphi \subseteq U$, then $V/U \cong V_1/U_1$.*

Proof: (1) For each subspace W of V with $Ker\ \varphi \subseteq W$, we put $W_1 = Im\ \varphi_W$, as in Theorem 21.1. Then W_1 is a subspace of V_1 by Theorem 41.8. and Example 41.6(d).

(2),(3),(4) These follow from parts (2),(3),(4) of Theorem 21.1 on regarding the subspaces merely as additive subgroups.

(5) From Theorem 21.1(5) and its proof, we know that $W := \{w \in V: w\varphi \in S\}$ is a subgroup of $(V,+)$ with $Ker\ \varphi \subseteq W$ and $W_1 = S$. For any $\alpha \in K$ and $w \in W$, we have $w\varphi \in S$, so $\alpha(w\varphi) \in S$, so $(\alpha w)\varphi \in S$, so $\alpha w \in W$ and W is in fact a subspace of V.

(6) Let $\nu': V_1 \to V_1/U_1$ be the natural homomorphism. Then ν' and $\varphi\nu': V \to V_1 \to V_1/U_1$ are vector space homomorphisms (Theorem 41.11, Theorem 41.7) with $Ker\ \varphi\nu' = U$ and $Im\ \varphi\nu' = V_1/U_1$ (Theorem 21.1(6),(7)). Hence, by Theorem 41.13, we have the vector space isomorphism

$$V/Ker\ \varphi\nu' \cong Im\ \varphi\nu'$$
$$V/U \cong V_1/U_1. \qquad \square$$

41.15 Theorem: *Let V be a vector space over a field K and let W be a subspace of V. The subspaces of V/W are given by U/W, where U runs through the subspaces of V containing W. In other words, for each subspace X of V/W, there is a unique subspace U of V such that $W \subseteq U$ and $X = U/W$. When X_1 and X_2 are subspaces of V/W, say with $X_1 = U_1/W$ and $X_2 = U_2/W$, where U_1, U_2 are subspaces of V containing W, then $X_1 \subseteq X_2$ if and only if $U_1 \subseteq U_2$. Furthermore, there holds*

$$V/W\ /\ U/W \cong V/U \qquad (\textit{vector space isomorphism}).$$

Proof: The natural homomorphism $\nu: V \to V/W$ is onto by Theorem 41.11. We may therefore apply Theorem 41.14. This theorem states that

any subspace of V/W is of the form $Im\ \nu_U$ for some subspace U of V with $Ker\ \nu \subseteq U$. Now

$$Im\ \nu_U = \{u\nu \in V/W: u \in U\}$$
$$= \{u + W \in V/W: u \in U\} = U/W$$

and $Ker\ \nu = W$ by Theorem 41.11. Thus the subspaces of V/W are given by U/W, where U's are subspaces of V containing W. By Theorem 41.14(2),(3),(4), $U_1/W \subseteq U_2/W$ if and only if $U_1 \subseteq U_2$, and $U_1/W \neq U_2/W$ whenever $U_1 \neq U_2$. Finally, by Theorem 41.14(6)

$$V/U \cong Im\ \nu_V/Im\ \nu_U = V/W\ /\ U/W \quad \text{as vector spaces.} \qquad \square$$

41.16 Theorem: *Let V be a vector space over a field K and let U,W be subspaces of V. Then $U \cap W$ and $U + W$ are subspaces of V and*

$$W/U \cap W \cong U + W/U \qquad \textit{(vector space isomorphism)}.$$

Proof: $U \cap W$ is a subspace of V by Example 40.6(g). Also, $U + W$ is a subgroup of $(V,+)$ by Lemma 19.4 and, for any $\alpha \in K$, $v \in U + W$, there are $u \in U$ and $w \in W$ with $v = u + w$, so that

$$\alpha v = \alpha(u + w) = \alpha u + \alpha w \in U + W$$

since $\alpha u \in U$ and $\alpha w \in W$; and so $U + W$ is closed under multiplication by scalars and $U + W$ is a subspace of V.

We consider the restriction

$$\nu_W: W \to V/U$$

to W of the natural homomorphism $\nu: V \to V/U$. By Theorem 41.11, ν is a vector space homomorphism; by Example 41.6(d), ν_W is a vector space homomorphism, so

$$W/Ker\ \nu_W \cong Im\ \nu_W \qquad \text{(as vector spaces)}$$

according to Theorem 41.13. From the proof of Theorem 21.3, we know that $Ker\ \nu_W = U \cap W$ and $Im\ \nu_W = U + W/U$, as may also be established directly. Hence

$$W/U \cap W \cong U + W/U. \qquad \square$$

Exercises

1. Let V be a vector space over a field K and let W be a subgroup of the additive group $(V,+)$. For all α in K and for all $v + W$ in the factor group

V/W, we write $\alpha \circ (v + W) = \alpha v + W$. Prove that $(\alpha, v + W) \to \alpha \circ (v + W)$ is a well defined mapping from $K \times (V/W)$ into V/W if and only if W is a subspace of V.

2. (cf. §20, Ex 14) Let $\varphi: V \to V_1$ be a vector space homomorphism, let W be a subspace of V such that $W \leqslant Ker\ \varphi$, and let $\nu: V \to V/W$ be the associated natural homomorphism. Show that there is a vector space homomorphism $\psi: V/W \to V_1$ such that $\nu\psi = \varphi$ and $Ker\ \psi = (Ker\ \varphi)/W$. What happens when we drop the condition $W \leqslant Ker\ \varphi$?

§42
Dependence and Bases

The span $s(A)$ of a subset A in vector space V is a subspace of V. This span may be the whole vector space V (we say then *A spans V*). In this paragraph, we study subsets A of V which span V and which are most economical in the sense that any proper subset of A spans a proper subspace of V.

We begin with a definition that will be important for everything in the sequel.

42.1 Definition: Let V be a vector space over a field K. A finite number of vectors $v_1, v_2, \ldots, v_n$ in V are called *linearly dependent over K* if there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$ in K, not all of them being zero, such that

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$$

(here 0 is the zero vector). If $v_1, v_2, \ldots, v_n$ are not linearly dependent over K, then $v_1, v_2, \ldots, v_n$ are said to be *linearly independent over K*.

A finite subset A of V is called *linearly dependent* (resp. *linearly independent*) *over K* if the finitely many vectors in A are linearly dependent (resp. linearly independent) over K.

An infinite subset A of V is called *linearly dependent over K* if there is a finite subset of A which is linearly dependent over K. An infinite subset A of V is called *linearly independent over K* if A is not linearly dependent over K, i.e., A is called linearly independent over K if every finite subset of A is linearly independent over K.

In place of the phrase "linearly (in)dependent over K", we shall also use the expression "K-linearly (in)dependent". When the field of scalars is clear from the context, we drop the phrase "over K" or the prefix "K-".

According to our definition, the vectors $v_1, v_2, \ldots, v_n$ of a vector space over K are linearly independent over K provided

$\alpha_1,\alpha_2,\ldots,\alpha_n \in K$, $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0 \implies \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$.
That is to say, $v_1,v_2,\ldots,v_n$ are K-linearly independent if the vector zero can be written as a linear combination of $v_1,v_2,\ldots,v_n$ only in the trivial way where the scalars are zero.

42.2 Examples: (a) Let V be a vector space over a field K and let v be a nonzero vector in V. Then $\alpha v = 0$ implies $\alpha = 0$ (Lemma 39.4(10)). Hence v (and $\{v\}$) is linearly independent over K. On the other hand, $\{0\}$ is linearly dependent over K because $1.0 = 0$ and $1 \neq 0$.

(b) Consider the vector space $\mathbb{Q}^3$ over $\mathbb{Q}$. The vectors $u = (1,0,0)$, $v = (0,1,0)$, $w = (0,0,1)$ of $\mathbb{Q}^3$ are linearly independent over $\mathbb{Q}$, for if $\alpha,\beta,\gamma \in K$ and $\alpha u + \beta v + \gamma w = 0$, then

$$\begin{aligned} \alpha(1,0,0) + \beta(0,1,0) + \gamma(0,0,1) &= (0,0,0) \\ (\alpha,0,0) + (0,\beta,0) + (0,0,\gamma) &= (0,0,0) \\ (\alpha,\beta,\gamma) &= (0,0,0) \\ \alpha = \beta = \gamma &= 0. \end{aligned}$$

(c) More generally, the vectors $u_1 = (1,0,\ldots,0)$, $u_2 = (0,1,\ldots,0)$, $\ldots,u_n = (0,0,\ldots,1)$ in the vector space K^n over a field K are linearly independent over K: if $\alpha_1,\alpha_2,\ldots,\alpha_n \in K$ and $\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n = 0$, then

$$\begin{aligned} \alpha_1(1,0,\ldots,0) + \alpha_2(0,1,\ldots,0) + \cdots + \alpha_n(0,0,\ldots,1) &= (0,0,\ldots,0) \\ (\alpha_1,0,\ldots,0) + (0,\alpha_2,\ldots,0) + \cdots + (0,0,\ldots,\alpha_n) &= (0,0,\ldots,0) \\ (\alpha_1,\alpha_2,\ldots,\alpha_n) &= (0,0,\ldots,0) \\ \alpha_1 = \alpha_2 = \cdots = \alpha_n &= 0. \end{aligned}$$

The reader is probably acquainted with the vectors u_1, u_2, u_3 in the vector space $\mathbb{R}^3$ over $\mathbb{R}$ under the names $\vec{i},\vec{j},\vec{k}$.

(d) The vectors $(1,0)$ and $(-1,0)$ in the $\mathbb{R}$-vector space $\mathbb{R}^2$ are linearly dependent over $\mathbb{R}$ because $1 \neq 0$ in $\mathbb{R}$ and $1(1,0) + 1(-1,0) = (0,0) =$ zero vector in $\mathbb{R}^2$.

(e) Let V be a vector space over a field K and let $v_1,v_2,\ldots,v_n$ be vectors in V which are linearly independent over K. Then any nonempty subset of $\{v_1,v_2,\ldots,v_n\}$ is linearly independent over K. In fact, if, say, $v_1,v_2,\ldots,v_m$ are linearly dependent over K ($m \leq n$), then there are scalars $\alpha_1,\alpha_2,\ldots,\alpha_m$ in K, not all equal to zero, such that

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_m v_m = 0;$$

hence, when we put (in case $m < n$) $\alpha_{m+1} = \cdots = \alpha_n = 0$, we obtain

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_m v_m + \alpha_{m+1} v_{m+1} + \cdots + \alpha_n v_n = 0,$$

where not all of $\alpha_1, \alpha_2, \ldots, \alpha_m, \alpha_{m+1}, \ldots, \alpha_n$ are equal to zero, contradicting the assumption that $v_1, v_2, \ldots, v_n$ are linearly independent over K. Thus any nonempty subset of a linearly independent finite set of vectors is linearly independent. But this statement is true also for infinite linearly independent sets. Indeed, let A be an infinite linearly independent subset of V and let B be a nonempty subset of A. If B is finite, then B is linearly independent by definition. If B is infinite, then any finite subset of B, being a finite subset of A, is linearly independent over K and hence B itself is linearly independent over K. Thus we have shown that *every nonempty subset of a linearly independent set of vectors is linearly independent*. Equivalently, *any set of vectors containing a linearly dependent subset is linearly dependent*.

(f) Let V be a vector space over a field K and let A be a subset of V containing $0 \in V$. Then A is linearly dependent over K by Example 42.2(a) and Example 42.2(e). Alternatively, just choose a finite number of vectors $v_1, v_2, \ldots, v_n$ from A including 0, say $v_1 = 0$ and observe that

$$1v_1 + 0v_2 + \ldots + 0v_n = 0,$$

so that $v_1, v_2, \ldots, v_n$ are linearly dependent over K and consequently A, too, is linearly dependent over K.

(g) Let V be the vector space $\mathbb{C}^2$ over $\mathbb{C}$. The vectors $(1,0)$, $(-i,0)$ in V are linearly dependent over $\mathbb{C}$, because

$$i(1,0) + 1(-i,0) = (0,0) = \text{zero vector in } V.$$

However, when V is regarded as an $\mathbb{R}$-vector space, these two vectors are not linearly dependent: if $\alpha, \beta \in \mathbb{R}$ and $\alpha(1,0) + \beta(-i,0) = (0,0)$, then $(\alpha - \beta i, 0) = (0,0)$, hence the complex number $\alpha - \beta i$ is equal to 0, so $\alpha = \beta = 0$. Thus $(1,0)$, $(-i,0)$ are linearly dependent over $\mathbb{C}$, but linearly independent over $\mathbb{R}$. This example shows that the field of scalars must be specified (unless it is clear from the context) whenever one discusses linear (in)dependence of vectors.

(h) Let V be a vector space over a field K and let $v_1, v_2, \ldots$ be infinitely many vectors in V. The linear dependence of $v_1, v_2, \ldots$ does *not* mean that there are scalars $\alpha_1, \alpha_2, \ldots$, not all equal to zero, such that

$$\sum_{k=1}^{\infty} \alpha_k v_k = 0.$$

This equation is meaningless, for its left hand side is not defined. What is defined (Definition 8.4) is a sum $\sum_{k=1}^{n} \alpha_k v_k$ of a *finite* number n of vectors $v_1, v_2, \ldots, v_n$ in V. The definition of $\sum_{k=1}^{\infty} \alpha_k v_k$ would involve some limiting process, and this is not possible in an arbitrary vector space.

(i) Consider the vector space $C^1([0,1])$ over $\mathbb{R}$ (Example 40.6(j)). The functions $f\colon [0,1] \to \mathbb{R}$ and $g\colon [0,1] \to \mathbb{R}$, where $f(x) = e^x$ and $g(x) = e^{2x}$ for all x in $[0,1]$, are vectors in $C^1([0,1])$. We claim that f and g are linearly independent over $\mathbb{R}$. To prove this, let us assume $\alpha, \beta \in \mathbb{R}$ and $\alpha f + \beta g =$ zero vector in $C^1([0,1])$. The zero vector in $C^1([0,1])$ is the function $z\colon [0,1] \to \mathbb{R}$ such that $z(x) = 0$ for all x in $[0,1]$. Hence

$$\begin{aligned} (\alpha f + \beta g)(x) &= 0 \qquad \text{for all } x \in [0,1], \\ \alpha f(x) + \beta g(x) &= 0 \qquad \text{for all } x \in [0,1], \\ \alpha e^x + \beta e^{2x} &= 0 \qquad \text{for all } x \in [0,1]. \end{aligned}$$

Differentiating, we obtain

$$\alpha e^x + 2\beta e^{2x} = 0 \qquad \text{for all } x \in [0,1].$$

We have thus $\beta e^{2x} = -\alpha e^x - 2\beta e^{2x}$ for all $x \in [0,1]$, hence $\beta = 0$, so $\alpha = 0$. Therefore f and g are linearly independent over $\mathbb{R}$.

(j) Let V be a vector space over a field K and let v_1, v_2 be vectors in V which are linearly dependent over K. Then there are scalars $\alpha, \beta \in K$, not both zero, such that $\alpha v_1 + \beta v_2 = 0$. If, say, $\alpha \neq 0$, then α has an inverse α^{-1} in K and we obtain $v_1 + (\alpha^{-1}\beta)v_2 = \alpha^{-1}(\alpha v_1 + \beta v_2) = \alpha^{-1}0 = 0$, so

$$v_1 = \gamma v_2$$

if we put $\gamma = -\alpha^{-1}\beta$. So v_1 is a scalar multiple of v_2. Conversely, if v_1 and v_2 are vectors in V and if one of them is a scalar multiple of the other, for instance if $v_1 = \gamma v_2$ with some $\gamma \in K$, then $1v_1 + (-\gamma)v_2 = 0$ and v_1, v_2 are linearly dependent over K. Thus the linear dependence of two vectors means that one of them is a scalar multiple of the other.

We generalize the last example.

42.3 Lemma: *Let V be a vector space over a field K and let $v_1, v_2, \ldots, v_n$ be n vectors in V, where $n \geqslant 2$. These vectors are linearly dependent over K if and only if one of them is a K-linear combination of the other vectors.*

Proof: We first assume that $v_1, v_2, \ldots, v_n$ are linearly dependent over K. Then there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$, not all of them zero, such that

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0.$$

To fix the ideas, let us suppose $\alpha_1 \neq 0$. Then α_1 has an inverse α_1^{-1} in K and we obtain $\alpha_1^{-1}(\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n) = \alpha_1^{-1} 0 = 0$,

$$v_1 + \alpha_1^{-1}\alpha_2 v_2 + \cdots + \alpha_1^{-1}\alpha_n v_n = 0,$$

$$v_1 = \gamma_2 v_2 + \cdots + \gamma_n v_n$$

where we put $\gamma_j = -\alpha_1^{-1}\alpha_j \in K$ $(j = 2, \ldots, n)$. So v_1 is a K-linear combination of the vectors $v_2, \ldots, v_n$.

Conversely, let us suppose that one of the vectors, for example v_1, is a linear combination of the rest, so that there are scalars $\alpha_2, \ldots, \alpha_n \in K$ such that

$$v_1 = \alpha_2 v_2 + \cdots + \alpha_n v_n.$$

Then we get

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$$

when we write $\alpha_1 = -1$. Since $\alpha_1 = -1 \neq 0$, we see that $v_1, v_2, \ldots, v_n$ are linearly dependent over K. □

A vector space over a field K can be spanned by many subsets of V. Among the subsets of V which span V, we want to find the ones with the least number of elements. The next two theorems, which are converses of each other, tell us that linearly dependent subsets are not useful for this purpose.

42.4 Theorem: *Let V be a vector space over a field K and let A be a nonempty subset of V. If A is linearly dependent over K, then there is a proper subset B of A such that $s_K(A) = s_K(B)$.*

Proof: Suppose A is K-linearly dependent. If A is infinite, then, by definition, there is a finite linearly dependent subset A_0 of A. If A is finite, let us put $A_0 = A$. Hence, in both cases, A_0 is a finite linearly dependent subset of A. Let $A_0 = \{v_0, v_1, v_2, \ldots, v_n\}$.

We first dispose of the trivial case $|A_0| = 1$, $V = A_0$. In this case we have $n = 0$ and $A_0 = \{v_0\}$, so $v_0 = 0$ by Example 42.2(a), so $A_0 = \{0\}$. Thus

$$\{0\} = A_0 \subseteq A \subseteq s_K(A) \subseteq V = A_0 = \{0\}$$

and $s_K(A)$ is equal to the K-span of the proper subset $B = \varnothing$ of A.

Suppose now $|A_0| \geq 1$ or $|A_0| = 1$ but $V \neq A_0$. Then we may and do join nonzero vectors $v_1, v_2, \ldots, v_n$ to A_0 without disturbing the linear dependence and finiteness of A_0. One of the vectors $v_0, v_1, v_2, \ldots, v_n$, which we may assume to be v_0 without loss of generality, is a K-linear combination of the others (Lemma 42.3). So there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$v_0 = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n.$$

We will show that v_0 is reduntant. We put $B = A \backslash \{v_0\}$. Then B is a proper subset of A and $s_K(B) \subseteq s_K(A)$. We prove $s_K(A) \subseteq s_K(B)$.

Let $v \in s_K(A)$. Then there are vectors $w_1, w_2, \ldots, w_m$ in A and scalars $\beta_1, \beta_2, \ldots, \beta_m$ in K such that

$$v = \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_m w_m.$$

Here we may suppose that $w_1, w_2, \ldots, w_m$ are pairwise distinct (if $w_1 = w_2$, we write $(\beta_1 + \beta_2)w_1$ instead of $\beta_1 w_1 + \beta_2 w_2$, etc.).

If none of the vectors $w_1, w_2, \ldots, w_m$ is equal to v_0, then v is a K-linear combination of the vectors $w_1, w_2, \ldots, w_m$ in B, so $v \in s_K(B)$.

If one of the vectors $w_1, w_2, \ldots, w_m$ is equal to v_0, for instance if $w_1 = v_0$, then we have

$$\begin{aligned} v &= \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_m w_m \\ &= \beta_1(\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n) + \beta_2 w_2 + \cdots + \beta_m w_m \\ &= \beta_1 \alpha_1 v_1 + \beta_1 \alpha_2 v_2 + \cdots + \beta_1 \alpha_n v_n + \beta_2 w_2 + \cdots + \beta_m w_m, \end{aligned}$$

so v is a K-linear combination of the vectors $v_1, v_2, \ldots, v_n, w_2, \ldots, w_m$ in $B = A \backslash \{v_0\}$ (some v_i might equal a w_j, but this does not matter), so $v \in s_K(B)$.

In both cases, $v \in s_K(B)$. Thus $s_K(A) \subseteq s_K(B)$ and $s_K(A) = s_K(B)$, as was to be proved. □

42.5 Theorem: *Let V be a vector space over a field K and let A be a nonempty subset of V. If there is a proper subset B of A such that $s_K(B) = s_K(A)$, then A is linearly dependent over K.*

Proof: We first dispose of the trivial case $B = \emptyset$. If $B = \emptyset$, then

$$\emptyset \neq A \subseteq s_K(A) = s_K(B) = s_K(\emptyset) = \{0\}$$

gives $A = \{0\}$ and A is K-linearly dependent by Example 42.2(a).

Suppose now $B \neq \emptyset$. Since $B \subset A$, there is a vector v in $A \backslash B$. From

$v \in A \subseteq s_K(A) = s_K(B)$, we conclude that there are vectors $w_1, w_2, \ldots, w_m$ in B and scalars $\beta_1, \beta_2, \ldots, \beta_m$ in K with

$$v = \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_m w_m.$$

So the vector v in A is a K-linear combination of the vectors $w_1, w_2, \ldots, w_m$ and the subset $\{v, w_1, w_2, \ldots, w_m\}$ of A is K-linearly dependent by Lemma 42.3. From Example 42.2(e), it follows that A is K-linearly dependent. □

The last two theorems lead us to consider linearly independent subsets of V spanning V. Whether an arbitrary vector space does have such a subset will be discussed later. We give a name to the subsets in question.

42.6 Definition: Let V be a vector space over a field K. A nonempty subset B of V is called a *basis of V over K*, or a *K-basis of V*, if B is linearly independent over K and spans V over K (i.e., $s_K(B) = V$). By convention, the empty set $\varnothing$ will be called a K-basis of the vector space $\{0\}$.

42.7 Examples: (a) Consider the vector space K^n over a field K. The vectors $u_1 = (1,0, \ldots,0)$, $u_2 = (0,1, \ldots,0)$, $\ldots$, $u_n = (0,0, \ldots,1)$ are linearly independent over K (Example 42.2(c)). Moreover, $\{u_1, u_2, \ldots, u_n\}$ spans K^n over K because any vector $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ in K^n is a K-linear combination

$$\alpha_1 u_1 + \alpha_2 u_2 + \ldots + \alpha_n u_n$$

of the vectors $u_1, u_2, \ldots, u_n$. Hence $\{u_1, u_2, \ldots, u_n\}$ is a basis of K^n over K.

(b) Let $V = \{h \in C^2([0,1]): h''(x) - 3h'(x) + 2h(x) = 0 \text{ for all } x \in [0,1]\}$. Then V is an $\mathbb{R}$-subspace of $C^2([0,1])$, as can be verified directly and also follows from Example 40.6(k). From the theory of ordinary differential equations, it is known that every function in V (that is, every solution of $y'' - 3y' + 2 = 0$) can be written in the form $c_1 f + c_2 g$, where $c_1, c_2 \in \mathbb{R}$ and $f(x) = e^x$, $g(x) = e^{2x}$ for all $x \in [0,1]$. Thus $\{f,g\}$ spans V over $\mathbb{R}$. Also, $\{f,g\}$ is linearly independent over $\mathbb{R}$ by Example 42.2(i). Hence $\{f,g\}$ is an $\mathbb{R}$-basis of V.

42.8 Theorem: *Let V be a vector space over a field K and let B*

$= \{v_1, v_2, \ldots, v_n\}$ *be a nonempty subset of* V. *Then* B *is a* K*-basis of* V *if and only if every element of* V *can be written in the form*

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n, \qquad (\alpha_1, \alpha_2, \ldots, \alpha_n \in K)$$

in a unique way (i.e., with unique scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$*).*

Proof: Assume first that B is a K-basis of V. Then $V = s_K(B)$, and every element of V can be written as

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$$

with suitable scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$. We are to show the uniqueness of this representation. In other words, we must prove that, if

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_n v_n \tag{1}$$

then $\alpha_1 = \beta_1, \alpha_2 = \beta_2, \ldots, \alpha_n = \beta_n$. This is easy: if (1) holds, then

$$(\alpha_1 - \beta_1)v_1 + (\alpha_2 - \beta_2)v_2 + \cdots + (\alpha_n - \beta_n)v_n = 0$$

and we obtain, since $B = \{v_1, v_2, \ldots, v_n\}$ is K-linearly independent, that $\alpha_1 - \beta_1 = \alpha_2 - \beta_2 = \cdots = \alpha_n - \beta_n = 0$. This proves uniqueness.

Conversely, let us suppose that every vector in V can be written in the form $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$ with unique scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$. Then $V = s_K(v_1, v_2, \ldots, v_n) = s_K(B)$. Moreover, B is linearly independent over K, for if $\alpha_1, \alpha_2, \ldots, \alpha_n \in K$ are scalars such that

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0,$$

then $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0v_1 + 0v_2 + \cdots + 0v_n$ and the uniqueness of the scalars in the representation of $0 \in V$ as a K-linear combination of $v_1, v_2, \ldots, v_n$ implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. Thus B is K-linearly independent and consequently B is a K-basis of V. □

We prove next that any finitely spanned vector space has a basis.

42.9 Theorem: *Let* V *be a vector space over a field* K *and assume* T *is a finite subset of* V *spanning* V, *so that* $s_K(T) = V$. *Then* V *has a finite* K*-basis. In fact, a suitable subset of* T *is a* K*-basis of* V.

Proof: If V happens to be the vector space $\{0\}$, then V has a K-basis, namely the empty set $\varnothing$ (Definition 42.6) and $\varnothing \subseteq T$. Having disposed of

this degenerate case, let us assume $V \neq 0$. Now $s_K(T) = V$. Since $V \neq 0$, we have $T \neq \emptyset$. If T is linearly independent over K, then T is a K-basis of V. Otherwise, there is a proper subset T_1 of T with $s_K(T_1) = s_K(T) = V$ (Theorem 42.4). Here $T_1 \neq \emptyset$, because $s_K(T_1) = V \neq \{0\}$. If T_1 is linearly independent over K, then T_1 is a K-basis of V. Otherwise, there is a proper subset T_2 of T_1 with $s_K(T_2) = s_K(T_1) = V$. Here $T_2 \neq \emptyset$, because $s_K(T_2) = V \neq \{0\}$. If T_2 is linearly independent over K, then T_2 is a K-basis of V. Otherwise, there is a proper subset T_3 of T_2 with $s_K(T_3) = s_K(T_2) = V$. Here $T_3 \neq \emptyset$, because $s_K(T_3) = V \neq \{0\}$. We continue in this way. Each time, we get a nonempty subset T_{i+1} of T_i such that $s_K(T_{i+1}) = V$ and T_{i+1} has less elements than T_i. Since T is a finite set, this process cannot go on indefinitely. Sooner or later, we will meet a K-linearly independent subset T_m of T with $s_K(T_m) = V$. This T_m is therefore a K-basis of V, and of course T_m is finite. □

Having convinced ourselves of the existence of bases in some vector spaces, we turn our attention to the number of vectors in a finite basis. We show that the number of linearly independent vectors in a subspace cannot exceed the number of vectors spanning the subspace. This theorem, due to E. Steinitz (1871-1928), is the source of many deep results concerning the dimension of a vector space. The idea is to replace some vectors in the spanning set by the vectors in the linearly independent set without changing the span.

42.10 Theorem (Steinitz' replacement theorem): *Let V be a vector space over a field K and $w_1, w_2, \ldots, w_m$ be finitely many vectors in V. Let $v_1, v_2, \ldots, v_n$ be n linearly independent vectors in the K-span $s_K(w_1, w_2, \ldots, w_m)$ of $w_1, w_2, \ldots, w_m$.*
Then $n \leq m$. Moreover, there are n vectors among $w_1, w_2, \ldots, w_m$, which we may assume to be $w_1, w_2, \ldots, w_n$, such that

$$s_K(v_1, v_2, \ldots, v_n, w_{n+1}, \ldots, w_m) = s_K(w_1, w_2, \ldots, w_m).$$

Proof: For $1 \leq h \leq n$, let A_h be the assertion

"there are h vectors among $w_1, w_2, \ldots, w_m$, say $w_1, \ldots, w_h$, such that

$$s_K(v_1, \ldots, v_h, w_{h+1}, \ldots, w_m) = s_K(w_1, \ldots, w_h, w_{h+1}, \ldots, w_m)".$$

We show that (1) A_1 is true,
(2) if $2 \leqslant h \leqslant n$ and A_{h-1} is true, then A_h is true.

This will establish $A_1, A_2, \ldots, A_{n-1}, A_n$. The second claim A_n in the theorem will be proved in this way.

(1) A_1 is true. We have $v_1 \in s_K(w_1, w_2, \ldots, w_m)$, so

$$v_1 = \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_m w_m,$$

with some scalars $\beta_1, \beta_2, \ldots, \beta_m \in K$. Since $v_1, v_2, \ldots, v_n$ are linearly independent over K, $v_1 \neq 0$ (Example 42.2(f)), hence not all of $\beta_1, \beta_2, \ldots, \beta_m$ are equal to $0 \in K$. So one of them is distinct from 0. Renaming $w_1, w_2, \ldots, w_m$ if necessary, we may suppose $\beta_1 \neq 0$. Then β_1 has an inverse β_1^{-1} in K and we get

$$w_1 = \beta_1^{-1}(v_1 - \beta_2 w_2 - \cdots - \beta_m w_m)$$
$$w_1 \in s_K(v_1, w_2, \ldots, w_m)$$
$$\{w_1, w_2, \ldots, w_m\} \subseteq s_K(v_1, w_2, \ldots, w_m). \qquad \text{(i)}$$

Since $v_1 \in s_K(w_1, w_2, \ldots, w_m)$,

we also have $\{v_1, w_2, \ldots, w_m\} \subseteq s_K(w_1, w_2, \ldots, w_m)$. (ii)

Using (i) and (ii) and applying Lemma 40.11 with $A = \{w_1, w_2, \ldots, w_m\}$ and $B = \{v_1, w_2, \ldots, w_m\}$, we obtain

$$s_K(v_1, w_2, \ldots, w_m) = s_K(w_1, w_2, \ldots, w_m).$$

This proves A_1.

(2) Suppose $2 \leqslant h \leqslant n$ and A_{h-1} is true. Then A_h is true. The truth of A_{h-1} means

$$s_K(v_1, \ldots, v_{h-1}, w_h, \ldots, w_m) = s_K(w_1, w_2, \ldots, w_m)$$

provided the w_i's are indexed suitably. We have

$$v_h \in s_K(w_1, \ldots, w_{h-1}, w_h, \ldots, w_m)$$
$$v_h \in s_K(v_1, \ldots, v_{h-1}, w_h, \ldots, w_m),$$

so

$$v_h = \alpha_1 v_1 + \cdots + \alpha_{h-1} v_{h-1} + \alpha_h w_h + \cdots + \alpha_m w_m$$

for some appropriate $\alpha_1, \ldots, \alpha_{h-1}, \alpha_h, \ldots, \alpha_m \in K$. Here not all of $\alpha_h, \ldots, \alpha_m$ are equal to $0 \in K$, for then v_h would be a K-linear combination $\alpha_1 v_1 + \cdots + \alpha_{h-1} v_{h-1}$ of the vectors $v_1, \ldots, v_{h-1}$ and the vectors $v_1, \ldots, v_{h-1}, v_h$ would not be linearly independent over K (Lemma 42.3), so $v_1, v_2, \ldots, v_n$ would not be linearly independent over K (Example 42.2(e)), contrary to the hypothesis. So one of $\alpha_h, \ldots, \alpha_m$ is distinct from 0. Renaming $w_h, \ldots, w_m$ if necessary, we may suppose $\alpha_h \neq 0$. Then α_h has an inverse α_h^{-1} in K and we get

$$-\alpha_h w_h = \alpha_1 v_1 + \cdots + \alpha_{h-1} v_{h-1} - v_h + \alpha_{h+1} w_{h+1} + \cdots + \alpha_m w_m,$$
$$w_h = -\alpha_h^{-1}(\alpha_1 v_1 + \cdots + \alpha_{h-1} v_{h-1} - v_h + \alpha_{h+1} w_{h+1} + \cdots + \alpha_m w_m),$$

$$w_h \in s_K(v_1, \ldots, v_{h-1}, v_h, w_{h+1}, \ldots, w_m).$$

Now each one of the vectors $w_1, \ldots, w_{h-1}$, being an element of the span $s_K(w_1, w_2, \ldots, w_m) = s_K(v_1, \ldots, v_{h-1}, w_h, \ldots, w_m)$, can be written in the form

$$\gamma_1 v_1 + \cdots + \gamma_{h-1} v_{h-1} + \gamma_h w_h + \gamma_{h+1} w_{h+1} + \cdots + \gamma_m w_m$$

with scalars $\gamma_1, \ldots, \gamma_{h-1}, \gamma_h, \gamma_{h+1}, \ldots, \gamma_m \in K$. Thus each one of $w_1, \ldots, w_{h-1}$ can be written as

$$\begin{aligned} &\gamma_1 v_1 + \cdots + \gamma_{h-1} v_{h-1} \\ &+ \gamma_h(-\alpha_h^{-1}(\alpha_1 v_1 + \cdots + \alpha_{h-1} v_{h-1} - v_h + \alpha_{h+1} w_{h+1} + \cdots + \alpha_m w_m)) \\ &+ \gamma_{h+1} w_{h+1} + \cdots + \gamma_m w_m, \end{aligned}$$

and so $\{w_1, \ldots, w_{h-1}\} \subseteq s_K(v_1, \ldots, v_{h-1}, v_h, w_{h+1}, \ldots, w_m)$.

Therefore $\{w_1, \ldots, w_{h-1}, w_h, w_{h+1}, \ldots, w_m\} \subseteq s_K(v_1, \ldots, v_{h-1}, v_h, w_{h+1}, \ldots, w_m)$. (i′)

Since $v_1, \ldots, v_{h-1}, v_h \in s_K(w_1, \ldots, w_{h-1}, w_h, w_{h+1}, \ldots, w_m)$,

we also have

$$\{v_1, \ldots, v_{h-1}, v_h, w_{h+1}, \ldots, w_m\} \subseteq s_K(w_1, \ldots, w_{h-1}, w_h, w_{h+1}, \ldots, w_m). \quad \text{(ii′)}$$

Using (i′) and (ii′) and applying Lemma 40.11 with

$$A = \{w_1, \ldots, w_{h-1}, w_h, w_{h+1}, \ldots, w_m\}, B = \{v_1, \ldots, v_{h-1}, v_h, w_{h+1}, \ldots, w_m\},$$

we obtain

$$s_K(v_1, \ldots, v_{h-1}, v_h, w_{h+1}, \ldots, w_m) = s_K(w_1, w_2, \ldots, w_m).$$

Thus A_h is true.

As remarked earlier, this establishes the truth of $A_1, A_2, \ldots, A_{n-1}, A_n$. In particular, A_n is true, and the second statement in the enunciation is proved. Now it remains to establish $n \leqslant m$.

If we had $m < n$, then A_m would be true and we would get

$$s_K(v_1, v_2, \ldots, v_m) = s_K(w_1, w_2, \ldots, w_m).$$

Then $v_n \in s_K(w_1, w_2, \ldots, w_m)$

would give $v_n \in {}_K(v_1, v_2, \ldots, v_m)$,

contrary to the hypothesis that $v_1, v_2, \ldots, v_m, \ldots, v_n$ are linearly independent over K. So $m < n$ is impossible and necessarily $n \leqslant m$. This completes the proof. □

42.11 Theorem: *Let V be a vector space over a field K and assume that V has a finite K-basis. Then any two K-bases of V have the same number of elements.*

Proof: There is a finite K-basis of V by hypothesis, say B. Assume that B has exactly n vectors ($n \geqslant 0$). We prove that any K-basis B_1 of V has also n vectors in it.

If $n = 0$, then $B = \varnothing$ and $V = \{0\}$. Thus $\varnothing$ and $\{0\}$ are the only subsets of V and $B = \varnothing$ is the only K-basis of V. Then any K-basis of V has exactly 0 elements.

Suppose now $n \geqslant 1$ and let B_1 be any K-basis of V. First we show that B_1 cannot be infinite. Otherwise, B_1 would be an infinite K-linearly independent subset of V. Every finite subset of B_1 would be K-linearly independent by definition. Let $v_1, v_2, \ldots, v_n, v_{n+1}$ be $n + 1$ K-linearly independent vectors in B_1. These $n + 1$ vectors lie in the K-span $s_K(B)$ of B and B has n elements. Steinitz' replacement theorem gives $n + 1 \leqslant n$, which is absurd. Thus B_1 cannot be infinite.

We put $|B_1| = n_1$. Here $n_1 \neq 0$, because $n_1 = 0$ would imply $B_1 = \varnothing$ and $\varnothing \neq B \subseteq s_K(B) = V = s_K(B_1) = s_K(\varnothing) = \{0\}$, so $B = \{0\}$ and B would be linearly independent over K, contrary to the hypothesis that B is a K-basis of V. So $n_1 \in \mathbb{N}$.

B_1 is a K-linearly independent subset of V in $s_K(B)$, therefore $n_1 \leqslant n$ by Steinitz' replacement theorem. Likewise, B is a K-linearly independent subset of V in $s_K(B_1)$, so $n \leqslant n_1$. Therefore $n = n_1$, as was to be proved. □

42.12 Definition: Let V be a vector space over a field K. If V has a finite K-basis, the number of elements in any K-basis of V, which is the same for all K-bases of V by Theorem 42.11, is called the *dimension of V over K*, or the *K-dimension of V*. It is denoted as $dim_K V$ or as $dim\, V$. If V has no finite K-basis, then the K-dimension of V is defined to be infinity, and we write in this case $dim_K V = \infty$.

Thus $dim_K K^n = n$ (Example 42.7(a)) and $dim_{\mathbb{R}} V = 2$, where V is the $\mathbb{R}$-vector space of Example 42.7(b).

We frequently say that V is n-dimensional when $dim\, V = n$. A vector space is said to be finite dimensional if $dim\, V$ is a nonnegative integer and infinite dimensional if $dim\, V = \infty$. Notice that the dimension of the vector space $\{0\}$ is zero.

42.13 Lemma: *Let V be a vector space over a field K, let $dim_K V = n \in \mathbb{N}$ and let $v_1, v_2, \ldots, v_n$ be n vectors in V.*
(1) If $v_1, v_2, \ldots, v_n$ are linearly independent over K, then $s_K(v_1, v_2, \ldots, v_n) = V$.
(2) If $s_K(v_1, v_2, \ldots, v_n) = V$, then $v_1, v_2, \ldots, v_n$ are linearly independent over K.

Proof: (1) We are given $dim_K V = n \in \mathbb{N}$. Let $\{w_1, w_2, \ldots, w_n\}$ be a basis of V over K. If $v_1, v_2, \ldots, v_n$ are K-linearly independent vectors in $V = s_K(w_1, w_2, \ldots, w_n)$, then we obtain $s_K(v_1, v_2, \ldots, v_n) = s_K(w_1, w_2, \ldots, w_n)$ by Steinitz' replacement theorem.

(2) Suppose $s_K(v_1, v_2, \ldots, v_n) = V$. If $v_1, v_2, \ldots, v_n$ are not linearly independent over K, then there is a proper subset $T \subset \{v_1, v_2, \ldots, v_n\}$ of $\{v_1, v_2, \ldots, v_n\}$ with $s_K(T) = s_K(v_1, v_2, \ldots, v_n) = V$ (Theorem 42.4), and there is a K-basis B of V such that $B \subseteq T$ (Theorem 42.9). Using Theorem 42.11, we obtain the contradiction

$$n = dim_K V = |B| \leqslant |T| < n.$$

Hence $v_1, v_2, \ldots, v_n$ have to be linearly independent over K. □

Any finite set spanning a vector space can be stripped off to a basis of that vector space (Theorem 42.9). Similarly, any linearly independent subset of a vector space can be extended to a basis, as we show now.

42.14 Theorem: *Let V be an m-dimensional vector space over a field K, with $m \geqslant 1$. Let $n \geqslant 1$ and let $v_1, v_2, \ldots, v_n$ be n linearly independent vectors in V. Then there is a K-basis B of V such that $\{v_1, v_2, \ldots, v_n\} \subseteq B$.*

Proof: Let $\{w_1, w_2, \ldots, w_m\}$ be a K-basis of V. Then $v_1, v_2, \ldots, v_n$ are linearly independent vectors in $V = s_K(w_1, w_2, \ldots, w_m)$, and Steinitz' replacement theorem gives

$$s_K(v_1, v_2, \ldots, v_n, w_{n+1}, \ldots, w_m) = V \quad \text{and } n \leqslant m$$

on indexing w's suitably. Then the $m = dim_K V$ vectors

$$v_1, v_2, \ldots, v_n, w_{n+1}, \ldots, w_m$$

are linearly independent over K by Lemma 42.13(2). Hence

$$B = \{v_1, v_2, \ldots, v_n, w_{n+1}, \ldots, w_m\}$$

is a K-basis of V containing the vectors $v_1, v_2, \ldots, v_n$. □

42.15 Lemma: *Let V be a finite dimensional vector space over a field K and let W be a subspace of V.*
(1) W is finite dimensional; in fact $dim_K W \leqslant dim_K V$.
(2) $dim_K W = dim_K V$ if and only if $W = V$.

Proof: Let $n = dim_K V$.

(1) The assertion is trivial when $W = \{0\}$, so let us assume $W \neq \{0\}$. Then there is a nonzero vector w in W, and $\{w\}$ is a K-linearly independent subset with one element. On the other hand, any $n + 1$ vectors in W (in fact in V) are linearly dependent over K by Steinitz' replacement theorem. Therefore there exists a natural number m such that

(a) $1 \leqslant m \leqslant n$,
(b) there are m linearly independent vectors in W,
(c) any $m + 1$ vectors in W are linearly dependent.

This m is clearly unique in view of (b) and (c). The natural number m having been defined in this way, let $w_1, w_2, \ldots, w_m$ be m K-linearly independent vectors in W. We claim that $\{w_1, w_2, \ldots, w_m\}$ is a K-basis of W. To show this, we must prove only that these vectors span W over K. Let w be an arbitrary vector in W. Then $w, w_1, w_2, \ldots, w_m$ are linearly dependent over K by (c). Hence

$$\beta w + \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_m w_m = 0$$

with some scalars $\beta, \beta_1, \beta_2, \ldots, \beta_m$ in K. Here $\beta \neq 0$, for otherwise the equation above would imply that $w_1, w_2, \ldots, w_m$ are K-linearly dependent. Hence β has an inverse β^{-1} in K and we get

$$w = (-\beta^{-1}\beta_1)w_1 + (-\beta^{-1}\beta_2)w_2 + \ldots + (-\beta^{-1}\beta_m)w_m,$$
$$w \in s_K(w_1, w_2, \ldots, w_m).$$

This gives $W \subseteq s_K(w_1, w_2, \ldots, w_m)$. But $w_1, w_2, \ldots, w_m$ belong to W, so $s_K(w_1, w_2, \ldots, w_m) \subseteq W$ (Lemma 40.5). The vectors $w_1, w_2, \ldots, w_m$ therefore span W over K, so $\{w_1, w_2, \ldots, w_m\}$ is a K-basis of W. Thus W is finite dimensional and in fact $dim_K W = m \leqslant n = dim_K V$.

(2) If $W = V$, then of course $dim_K W = dim_K V$. Suppose conversely $dim_K W = dim_K V = n$ and let A be a K-basis of W. Then there is a K-basis B of V with $A \subseteq B$: this follows from Theorem 42.14 when $A \neq \varnothing$ and is obvious when $A = \varnothing$. Then

$$n = dim_K W = |A| \leqslant |B| = dim_K V = n$$

implies that $A = B$. Thus $W = s_K(A) = s_K(B) = V$. □

42.16 Lemma: *Let V, U be vector spaces over a field K. Suppose V is finite dimensional and let $\varphi: V \to U$ be a vector space homomorphism. Let $v_1, v_2, \ldots, v_n$ be vectors in V.*
(1) If φ is one-to-one and $\{v_1, v_2, \ldots, v_n\}$ is linearly independent over K, then $\{v_1\varphi, v_2\varphi, \ldots, v_n\varphi\}$ is linearly independent over K.
(2) If φ is onto U and $\{v_1, v_2, \ldots, v_n\}$ spans V over K, then $\{v_1\varphi, v_2\varphi, \ldots, v_n\varphi\}$ spans U over K.
(3) If φ is a vector space isomorphism and $\{v_1, v_2, \ldots, v_n\}$ is a K-basis of V, then $\{v_1\varphi, v_2\varphi, \ldots, v_n\varphi\}$ is a K-basis of U. In particular, $dim_K U = dim_K V$.

Proof: (1) Suppose $\alpha_1, \alpha_2, \ldots, \alpha_n$ are scalars such that

$$\alpha_1(v_1\varphi) + \alpha_2(v_2\varphi) + \ldots + \alpha_n(v_n\varphi) = 0.$$

Then $$(\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n)\varphi = 0,$$

so $$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n \in Ker\ \varphi,$$

and $$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$$

since $Ker\ \varphi = 0$ as φ is one-to-one. Since $v_1, v_2, \ldots, v_n$ are K-linearly independent, we get $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. Hence $v_1\varphi, v_2\varphi, \ldots, v_n\varphi$ are linearly independent over K.

(2) We must show that any element of U can be written as a K-linear combination of the vectors $v_1\varphi, v_2\varphi, \ldots, v_n\varphi$. Let $u \in U$. Then there is a v in V with $v\varphi = u$ and $v = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$, where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are suitable scalars in K. This yields

$$\begin{aligned} u = v\varphi &= (\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n)\varphi \\ &= \alpha_1(v_1\varphi) + \alpha_2(v_2\varphi) + \cdots + \alpha_n(v_n\varphi) \in s_K(v_1\varphi, v_2\varphi, \ldots, v_n\varphi). \end{aligned}$$

as was to be proved.

(3) This follows immediately from (1) and (2). □

From Lemma 42.16, it follows that $dim_K U = n$ whenever $U \cong K^n$. The converse of this statement is also true.

42.17 Theorem: *Let V be a vector space over a field K. Then $dim_K V = n \in \mathbb{N}$ if and only if $V \cong K^n$* (as vector spaces).

Proof: If $V \cong K^n$, then $dim_K V = dim_K K^n = n$ by Lemma 42.16(3). Suppose conversely that $dim_K V = n$ and let $\{v_1, v_2, \ldots, v_n\}$ be a K-basis of V. Every element v of V can be written in a unique way as $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$, where $\alpha_1, \alpha_2, \ldots, \alpha_n \in K$. We consider the mapping

$$\varphi: V \longrightarrow K^n$$
$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n \longrightarrow (\alpha_1, \alpha_2, \ldots, \alpha_n).$$

This φ is a K-linear transformation, since, for any $\alpha, \beta \in K$ and $v = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$, $w = \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_n v_n$ in V, we have

$$\begin{aligned}(\alpha v + \beta w)\varphi &= (\alpha(\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n) + \beta(\beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_n v_n))\varphi \\ &= ((\alpha\alpha_1 + \beta\beta_1)v_1 + (\alpha\alpha_2 + \beta\beta_2)v_2 + \cdots + (\alpha\alpha_n + \beta\beta_n)v_n)\varphi \\ &= (\alpha\alpha_1 + \beta\beta_1, \alpha\alpha_2 + \beta\beta_2, \ldots, \alpha\alpha_n + \beta\beta_n) \\ &= \alpha(\alpha_1, \alpha_2, \ldots, \alpha_n) + \beta(\beta_1, \beta_2, \ldots, \beta_n) \\ &= \alpha(v\varphi) + \beta(w\varphi).\end{aligned}$$

Furthermore, $v = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n \in V$ belongs to $Ker\ \varphi$ if and only if $(\alpha_1, \alpha_2, \ldots, \alpha_n) = (0,0, \ldots, 0)$, thus if and only if $v = 0v_1 + 0v_2 + \cdots + 0v_n = 0$. So $Ker\ \varphi = \{0\}$ and φ is one-to-one.

Since any n-tuple $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ in K^n is the image, under φ, of the vector $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$ in V, we see that φ is onto.

Hence φ is a vector space isomorphism and $V \cong K^n$. □

42.18 Theorem: *Let V and U be finite dimensional vector spaces over a field K. Then $V \cong U$ if and only if $dim_K V = dim_K U$.*

Proof: The case when $dim_K V = 0$ or $dim_K U = 0$ is trivial. Let us suppose $dim_K V \geq 1$ and $dim_K U \geq 1$. If $V \cong U$, then $dim_K V = dim_K U$ by Lemma 42.16(3). If $dim_K V = dim_K U$, then $V \cong K^{dim_K V} = K^{dim_K U} \cong U$ by Theorem 42.17, hence $V \cong U$. □

42.19 Theorem: *Let V be a vector space over a field K and let W be a subspace of V. If V is finite dimensional, then V/W is finite dimensional. In fact,*

$$dim_K V = dim_K W + dim_K V/W.$$

Proof: We eliminate the trivial cases. We know that W is finite dimensional (Lemma 42.15(1)). If $dim_K W = 0$, then $W = \{0\}$, so $V \cong V/\{0\} = V/W$, so $dim_K V/W = dim_K V$ and $dim_K V = 0 + dim_K V = dim_K W + dim_K V/W$. If $dim_K W = dim_K V$, then $W = V$ (Lemma 42.15(2)), so $V/W \cong \{0\}$ and $dim_K V = dim_K V + 0 = dim_K W + dim_K V/W$. Thus the theorem is proved in case $dim_K W = 0$ or $dim_K W = dim_K V$ (in particular in case $dim_K V = 0$).

Let us assume now $0 < dim_K W < dim_K V$. Let $dim_K W = m$ and let $\{w_1, w_2, \ldots, w_m\}$ be a K-basis of W. There are vectors $u_1, u_2, \ldots, u_k$ in V such that $\{w_1, w_2, \ldots, w_m, u_1, u_2, \ldots, u_k\}$ is a K-basis of V (Theorem 42.14). Here $k \geq 1$ and $m + k = dim_K V$. We claim that $\{u_1 + W, u_2 + W, \ldots, u_k + W\}$ is a K-basis of V/W. This will imply $k = dim_K V/W$, hence $dim_K V = m + k = dim_K W + dim_K V/W$.

To establish our claim, we note first that $u_1 + W, u_2 + W, \ldots, u_k + W$ are K-linearly independent vectors in V/W. Indeed, if $\alpha_1, \alpha_2, \ldots, \alpha_k$ are scalars such that

$$\alpha_1(u_1 + W) + \alpha_2(u_2 + W) + \cdots + \alpha_k(u_k + W) = 0 + W,$$

then

$$\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_k u_k \in W = s_K(w_1, w_2, \ldots, w_m),$$

$$\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_k u_k = \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_m w_m$$

where $\beta_1, \beta_2, \ldots, \beta_m$ are appropriate scalars in K. Then

$$\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_k u_k - \beta_1 w_1 - \beta_2 w_2 - \cdots - \beta_m w_m = 0$$

and linearly independence of $w_1, w_2, \ldots, w_m, u_1, u_2, \ldots, u_k$ implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$. Thus $u_1 + W, u_2 + W, \ldots, u_k + W$ in V/W are linearly independent over K.

Secondly, these vectors span V/W. To see this, let us take an arbitrary vector $v + W$ in V/W, where $v \in V$. Then

$$v = \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_k u_k + \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_m w_m$$

where $\alpha_1, \alpha_2, \ldots, \alpha_k, \beta_1, \beta_2, \ldots, \beta_m$ are scalars, and thus

$$\begin{aligned} v + W &= (\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_k u_k + \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_m w_m) + W \\ &= \alpha_1(u_1 + W) + \alpha_2(u_2 + W) + \cdots + \alpha_k(u_k + W) \\ &\quad + \beta_1(w_1 + W) + \beta_2(w_2 + W) + \cdots + \beta_m(w_m + W) \\ &= \alpha_1(u_1 + W) + \alpha_2(u_2 + W) + \cdots + \alpha_k(u_k + W) \\ &\in s_K(u_1 + W, u_2 + W, \ldots, u_k + W) \\ &\subseteq V/W, \end{aligned}$$

hence $V/W = s_K(u_1 + W, u_2 + W, \ldots, u_k + W)$. This proves that $\{u_1 + W, u_2 + W, \ldots, u_k + W\}$ is a basis of V/W over K. As we remarked above, this gives $dim_K V = dim_K W + dim_K V/W$. □

We deduce important corollaries from Theorem 42.19.

42.20 Theorem: *Let V be a vector space over a field K. Let W, U be finite dimensional subspaces of V. Then $W + U$ is a finite dimensional subspace of V and in fact*

$$dim_K(W + U) = dim_K W + dim_K U - dim_K(W \cap U).$$

Proof: If $\{w_1, w_2, \ldots, w_m\}$ is a K-basis of W and $\{u_1, u_2, \ldots, u_k\}$ is a K-basis of U, then $W + U = \{w + u \in V : w \in W, u \in U\}$ is clearly spanned by the finite set $\{w_1, w_2, \ldots, w_m, u_1, u_2, \ldots, u_k\}$, hence $W + U$ is finite dimensional (Theorem 42.9). From $W + U/U \cong W/W \cap U$ (Theorem 41.16), we obtain then

$$\begin{aligned} dim_K(W + U) - dim_K U &= dim_K(W + U/U) \\ &= dim_K(W/W \cap U) \\ &= dim_K W - dim_K(W \cap U). \end{aligned}$$

□

42.21 Theorem: *Let V be a vector space over a field K and let φ be a K-linear transformation from V. If V is finite dimensional, then*

$$dim_K Ker\, \varphi + dim_K Im\, \varphi = dim_K V.$$

Proof: Theorem 41.13 tells us $V/Ker\ \varphi \cong Im\ \varphi$ and Theorem 42.19 gives

$$dim_K V - dim_K Ker\ \varphi = dim_K Im\ \varphi.$$ □

42.22 Theorem: *Let V,U be vector spaces over a field K and let $\varphi: V \to U$ be a K-linear mapping. Suppose that V and U have the same finite dimension. Then the following statements are equivalent.*

(1) *φ is one-to-one.*

(2) *φ is onto.*

(3) *φ is a vector space isomorphism.*

Proof: (1) $\Longrightarrow$ (2) If φ is one-to-one, then $Ker\ \varphi = \{0\}$, so $dim_K Ker\ \varphi = 0$ and $dim_K Im\ \varphi = dim_K Ker\ \varphi + dim_K Im\ \varphi = dim_K V = dim_K U$. Thus $Im\ \varphi$ is a subspace of U with $dim_K Im\ \varphi = dim_K U$, and Lemma 42.15(2) gives then $Im\ \varphi = U$. Hence φ is onto.

(2) $\Longrightarrow$ (1) If φ is onto, then $Im\ \varphi = U$, so $dim_K Im\ \varphi = dim_K U$ and $dim_K Ker\ \varphi = dim_K V - dim_K Im\ \varphi = dim_K U - dim_K U = 0$. Thus $Ker\ \varphi = \{0\}$ and φ is one-to-one.

Hence any one of (1),(2) implies the other, and these together imply (3). Conversely, if φ is an isomorphism, then of course φ is one-to-one and onto. Thus (3) implies both (1) and (2). □

We close this paragraph with a brief discussion of infinite dimensional vector spaces. Do infinite dimensional vector spaces have bases? From Theorem 42.9, we know that such a vector space cannot be spanned by a finite set. But if B is a spanning set, necessarily infinite, the argument of Theorem 42.9 does not work. To prove the existence of bases of infinite dimensional vector spaces, we have to resort to more sophisticated means.

It is in fact true that every vector space has a basis, and a proof is given in the appendix. The proof of this statement for infinite dimensional vector spaces requires a fundamental tool known as Zorn's lemma. This lemma can be used in a variety of situations to establish the existence of certain objects.

The existence of bases having been assured by Zorn's lemma, we might ask whether any two bases have the same cardinality. The answer turned out to be "yes" in the finite dimensional case (Theorem 42.11), and this was proved by using Steinitz' replacement theorem. The proof of Steinitz' replacement theorem does not extend to the infinite dimensional case. Nevertheless, theorems of set theory can be employed to show that two bases of a vector space have the same cardinal number. Thus renders it possible to define the dimension of a vector space as the cardinality of a basis. Hence it is possible to distinguish between various types of infinities. This is much finer than Definition 42.12, by which infinite dimensionality is merely a crude negation of finite dimensionality.

Theorem 42.14, which states that any linearly independent subset can be extended to a basis, is true in the infinite dimensional case, too. The proof makes use of Zorn's lemma.

Lemma 42.15(1) remains valid also in the infinite dimensional case, in the sense that a basis of a subspace has a cardinal number less than or equal to the cardinality of a basis of the whole space. Lemma 42.15(2), however, is not necessarily true for infinite dimensional vector spaces: a proper subspace may have the same dimension as the whole space (think of $\mathbb{R}$ and $\mathbb{C}$ as $\mathbb{Q}$-vector spaces).

Lemma 42.16 and its proof works in the infinite dimensional case.

Lemma 42.19 and its proof works in the infinite dimensional case, provided we refer to the generalization of Theorem 42.14 at the appropriate place.

Generally speaking, infinite dimensional vector spaces are wild objects. To render them more manageable, one equips them with some additional structure, perhaps with a topological or analytic one.

Exercises

1. Let V be a vector space over a field K and let W be a subspace of V. Show that there is a subspace U of V such that $V = W + U$ and

$W \cap U = \{0\}$. (U is called a *direct complement of* W *in* V. We write then $V = W \oplus U$ and call V the *direct sum of* W *and* U.)

2. Is $\{(1,1,1), (1,1,0), (1,0,0)\}$ an $\mathbb{R}$-basis of $\mathbb{R}^3$?

3. Is $\{(\bar{1},\bar{2},\bar{6},), (\bar{0},\bar{0},\bar{1},), (\bar{2},\bar{1},\bar{0},)$ a $\mathbb{Z}_3$-basis of $\mathbb{Z}_3^3$?

4. Find an $\mathbb{R}$-basis of

$$\{f \in C^2([0,1]): f''(x) - 7f'(x) + 12f(x) = 0 \text{ for all } x \in [0,1]\}.$$

5. Find all $\mathbb{R}$-linear mappings from $\mathbb{R}^4$ onto $\mathbb{R}^5$.

6. Find all $\mathbb{Z}_2$-bases of $\mathbb{Z}_2^3$ and $\mathbb{Z}_3$-bases of $\mathbb{Z}_3^2$.

7. Show that the vectors $(1,2,1)$, $(0,2,0)$, $(1,2,-1)$ and also the vectors $(1,1,0)$, $(1,0,1)$, $(1,1,1)$ in $\mathbb{Q}^3$ are linearly independent over $\mathbb{Q}$.

8. Let $f_k(x) = \sin kx$ for $x \in [0,1]$ $(k = 1,2,3, \ldots)$. Prove that the functions $\{f_1, f_2, f_3, \ldots\}$ in $C^\infty([0,1])$ are linearly independent over $\mathbb{R}$.

§43
Linear Transformations and Matrices

In this paragraph, we learn to construct a new vector space from two given vector spaces V,W, namely the vector space of linear transformations from V into W. We introduce matrices and study the relationship between linear transformations and matrices.

Suppose V and W are vector spaces over a field K. We denote by $L_K(V,W)$ the set of all K-linear mappings from V into W. This set $L_K(V,W)$ is not empty, for at least the mapping $V \to W$ is a K-linear
$v \to 0$
transformation in $L_K(V,W)$. We want to define an addition and a multiplication by scalars on $L_K(V,W)$ and make $L_K(V,W)$ into a K-vector space.

Let $T,S \in L_K(V,W)$. How shall we define $T+S$? Well, the only natural way to define $T+S$ is to put $v(T+S) = vT + vS$ for all $v \in V$ (pointwise addition). What about multiplication by scalars? Given $\alpha \in K$ and $T \in L(V,W)$, the mapping αT had better mean: first multiply by α, then apply T, so that $v(\alpha T) := (\alpha v)T$ (or, first apply T, then multiply by α, so that $v(\alpha T) := \alpha(vT)$, but this is the same definition as before).

43.1 Theorem: *Let V,W be vector spaces over a field K and let $L_K(V,W)$ be the set of all K-linear transformations from V into W. For any T,S in $L_K(V,W)$ and for any α in K, we write*

$$v(T+S) = vT + vS, \qquad v(\alpha T) = (\alpha v)T \qquad (v \in V).$$

Under this addition and multiplication by scalars, $L_K(V,W)$ is a vector space over K.

Proof: We show first that $L_K(V,W)$ is an abelian group under addition.

(i) Let $T,S \in L_K(V,W)$. Then $(\alpha v_1 + \beta v_2)(T+S)$
$$= (\alpha v_1 + \beta v_2)T + (\alpha v_1 + \beta v_2)S$$
$$= \alpha(v_1T) + \beta(v_2T) + \alpha(v_1S) + \beta(v_2S)$$
$$= \alpha(v_1T) + \alpha(v_1S) + \beta(v_2T) + \beta(v_2S)$$

$$= \alpha(v_1T + v_1S) + \beta(v_2T + v_2S)$$
$$= \alpha(v_1(T + S)) + \beta(v_2(T + S))$$

for all $\alpha,\beta \in K$ and $v_1,v_2 \in V$. Thus $T + S$ is K-linear and $T + S \in \boldsymbol{L}_K(V,W)$. Therefore $\boldsymbol{L}_K(V,W)$ is closed under addition.

(ii) Let T,S,R be arbitrary elements of $\boldsymbol{L}_K(V,W)$. Then

$$v((T + S) + R) = v(T + S) + vR = (vT + vS) + vR$$
$$= vT + (vS + vR) = vT + v(S + R) = v(T + (S + R))$$

for all $v \in V$; hence $(T + S) + R = T + (S + R)$. Thus addition in $\boldsymbol{L}_K(V,W)$ is associative.

(iii) Let $0^*: V \to W$, $v \to 0$. Then $(\alpha v_1 + \beta v_2)0^* = 0 = \alpha 0 + \beta 0$

$= \alpha(v_1 0^*) + \beta(v_2 0^*)$ for all $\alpha,\beta \in K$, $v_1,v_2 \in V$ and 0^* is in $\boldsymbol{L}_K(V,W)$. From

$$v(T + 0^*) = vT + v0^* = vT + 0 = vT \qquad (v \in V),$$

we obtain $T + 0^* = T$ for any $T \in \boldsymbol{L}_K(V,W)$. Thus 0^* is a right identity.

(iv) Any $T \in \boldsymbol{L}_K(V,W)$ has an opposite in $\boldsymbol{L}_K(V,W)$, namely the mapping $S: V \to W$, $v \to -(vT)$. Indeed

$$v(T + S) = vT + vS = vT + (-(vT)) = 0 = v0^*$$

for all $v \in V$, so $T + S = 0^*$. Are we done? No! We should check that S is in fact in $\boldsymbol{L}(V,W)$, but this easy:

$$(\alpha v_1 + \beta v_2)S = ((\alpha v_1 + \beta v_2)T)$$
$$= (-(\alpha v_1 + \beta v_2))T \qquad \text{(Lemma 41.5(2))}$$
$$= ((-\alpha v_1) + (-\beta v_2))T$$
$$= (\alpha(-v_1) + \beta(-v_2))T$$
$$= \alpha((-v_1))T + \beta((-v_2))T$$
$$= \alpha(-(v_1T)) + \beta(-(v_2T))$$
$$= \alpha(v_1S) + \beta(v_2S)$$

for all $\alpha,\beta \in K$ and $v_1,v_2 \in V$. Thus S is in $\boldsymbol{L}_K(V,W)$ and S is a right inverse of T.

(v) Finally, $T + S = S + T$ for any $T,S \in \boldsymbol{L}_K(V,W)$, because

$$v(T + S) = vT + vS = vS + vT = v(S + T)$$

for all $v \in V$. Hence $\boldsymbol{L}_K(V,W)$ is a commutative group under addition.

Now the properties of multiplication by scalars. First we note that αT is in $\boldsymbol{L}(V,W)$ whenever $\alpha \in K$ and $T \in \boldsymbol{L}_K(V,W)$, because

$$(v_1 + v_2)(\alpha T) = (\alpha(v_1 + v_2))T = (\alpha v_1 + \alpha v_2)T$$
$$= (\alpha v_1)T + (\alpha v_2)T = v_1(\alpha T) + v_2(\alpha T)$$

for all $v_1, v_2 \in V$, so that αT is additive and

$$(\beta v)(\alpha T) = (\alpha(\beta v))T = ((\alpha\beta)v)T = ((\beta\alpha)v)T = ((\beta(\alpha v))T = \beta((\alpha v)T) = \beta(v(\alpha T))$$

for all $\beta \in K$, $v \in V$, so that αT is homogeneous. So αT belongs to $L_K(V,W)$.

(1) $\alpha(T + S) = \alpha T + \alpha S$ for all $\alpha \in K$ and $T, S \in L_K(V,W)$ since

$$v(\alpha(T + S)) = (\alpha v)(T + S) = (\alpha v)T + (\alpha v)S = v(\alpha T) + v(\alpha S) = v(\alpha T + \alpha S)$$

for any $v \in V$.

(2) $(\alpha + \beta)T = \alpha T + \beta T$ for all $\alpha, \beta \in K$ and $T \in L_K(V,W)$ since

$$v((\alpha + \beta)T) = ((\alpha + \beta)v)T = (\alpha v + \beta v)T$$
$$= (\alpha v)T + (\beta v)T = v(\alpha T) + v(\beta T) = v(\alpha T + \beta T)$$

for any $v \in V$.

(3) $(\alpha\beta)T = \alpha(\beta T)$ for all $\alpha, \beta \in K$ and $T \in L_K(V,W)$ since

$$v((\alpha\beta)T) = ((\alpha\beta)v)T = ((\beta\alpha)v)T = (\beta(\alpha v))T = (\alpha v)(\beta T) = v(\alpha(\beta T))$$

for any $v \in V$.

(4) $1T = T$ for all $T \in L_K(V,W)$ since

$$v(1T) = (1v)T = (v)T = vT$$

for any $v \in V$.

Thus $L_K(V,W)$ is a K-vector space. □

Let us assume now that V is an n-dimensional K-vector space and W is an m-dimensional K-vector space ($n, m \in \mathbb{N}$). Let $B = \{v_1, v_2, \ldots, v_n\}$ be a K-basis of V and $B' = \{w_1, w_2, \ldots, w_m\}$ be a K-basis of W. Then, for any linear transformation T in $L_K(V,W)$, we have

$$\begin{aligned} v_1 T &= \alpha_{11} w_1 + \alpha_{12} w_2 + \cdots + \alpha_{1m} w_m \\ v_2 T &= \alpha_{21} w_1 + \alpha_{22} w_2 + \cdots + \alpha_{2m} w_m \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ v_n T &= \alpha_{n1} w_1 + \alpha_{n2} w_2 + \cdots + \alpha_{nm} w_m \end{aligned} \qquad (*)$$

where $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{nm}$ are scalars in K. The arrangement of the scalars in (*) deserves a name.

43.2 Definition: Let K be a field and $n,m \in \mathbb{N}$. An *n by m matrix over K* is an array

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nm} \end{pmatrix} \tag{1}$$

of nm elements $\alpha_{11},\alpha_{12},\ldots,\alpha_{nm}$ of K, arranged in n rows and m columns, and enclosed within parentheses. The set of all n by m matrices over K will be denoted by $Mat_{n\times m}(K)$.

Sometimes we write "$n\times m$" instead of "n by m". The horizontal lines

$$\alpha_{i1}\ \alpha_{i2}\ \cdots\ \alpha_{im} \tag{2}$$

of a matrix over K are called the *rows* of that matrix. More specifically, (2) is the i-th row of the matrix (1). The vertical lines

$$\begin{matrix} \alpha_{1j} \\ \alpha_{2j} \\ \vdots \\ \alpha_{nj} \end{matrix} \tag{3}$$

of a matrix over K are called the *columns* of that matrix. More specifically, (3) is the j-th column of the matrix (1). The element α_{ij} is at the place where the i-th row and the j-th column meet. The first index i refers to the row, the second index j refers to the column. Also, in the expression "n by m", the first number n specifies the number of rows, the second number m specifies the number of columns of the matrix. The elements α_{ij} are called the *entries* of the matrix (1). When $n = m$, the matrix (1) is said to be a *square* matrix. The set of all square matrices with n rows (or n columns) over K will be denoted by $Mat_n(K)$ (instead of $Mat_{n\times n}(K)$).

We will usually abbreviate the matrix (1) as (α_{ij}).

Two matrices $(\alpha_{ij}) \in Mat_{n\times m}(K)$ and $(\beta_{ij}) \in Mat_{n'\times m'}(K)$ are declared to be *equal* if $n = n'$, $m = m'$ and $\alpha_{ij} = \beta_{ij}$ for all $i = 1,2,\ldots,n$; $j = 1,2,\ldots,m$. Thus two matrices are equal if and only if they have the same number of rows and columns, and have the same elements at corresponding places. We write then $(\alpha_{ij}) = (\beta_{ij})$. Otherwise, we put $(\alpha_{ij}) \neq (\beta_{ij})$.

We now make $Mat_{n\times m}(K)$ into a vector space over K.

43.3 Definition: Let K be a field, $\alpha \in K$ and let $A, B \in Mat_{n\times m}(K)$, say $A = (\alpha_{ij})$, $B = (\beta_{ij})$. We write

$A + B = C$, C being the matrix (γ_{ij}) in $Mat_{n\times m}(K)$, where $\gamma_{ij} = \alpha_{ij} + \beta_{ij}$, and $\alpha A = E$, E being the matrix (ε_{ij}) in $Mat_{n\times m}(K)$, where $\varepsilon_{ij} = \alpha\alpha_{ij}$. In other words, $(\alpha_{ij}) + (\beta_{ij}) = (\alpha_{ij} + \beta_{ij})$ and $\alpha(\alpha_{ij}) = (\alpha\alpha_{ij})$.

43.4 Theorem: *Let K be a field. Under the addition and multiplication by scalars of Definition* 43.3, *the set $Mat_{n\times m}(K)$ is a vector space over K.*

Proof: First we check that $Mat_{n\times m}(K)$ is an abelian group under addition.

(i) For any $A = (\alpha_{ij})$, $B = (\beta_{ij})$ in $Mat_{n\times m}(K)$, we have $A + B = (\alpha_{ij} + \beta_{ij})$ and each $\alpha_{ij} + \beta_{ij}$ is an element of K, because K is closed under addition ($i = 1,2, \ldots, n$; $j = 1,2, \ldots, m$). Hence $A + B \in Mat_{n\times m}(K)$ and $Mat_{n\times m}(K)$ is closed under addition.

(ii) For any $A = (\alpha_{ij})$, $B = (\beta_{ij})$, $C = (\gamma_{ij})$ in $Mat_{n\times m}(K)$, there holds

$$(A + B) + C = ((\alpha_{ij}) + (\beta_{ij})) + (\gamma_{ij}) = (\alpha_{ij} + \beta_{ij}) + (\gamma_{ij}) = ((\alpha_{ij} + \beta_{ij}) + \gamma_{ij})$$
$$= (\alpha_{ij} + (\beta_{ij} + \gamma_{ij})) = (\alpha_{ij}) + (\beta_{ij} + \gamma_{ij}) = (\alpha_{ij}) + ((\beta_{ij}) + (\gamma_{ij})) = A + (B + C)$$

and addition in $Mat_{n\times m}(K)$ is associative.

(iii) Let $\tilde{0}$ be the n by m matrix whose entries are all equal to the zero element of K. Thus $\tilde{0} = (\zeta_{ij})$, where $\zeta_{ij} = 0 \in K$ for all i,j. Then

$$A + \tilde{0} = (\alpha_{ij}) + (\zeta_{ij}) = (\alpha_{ij} + \zeta_{ij}) = (\alpha_{ij} + 0) = (\alpha_{ij}) = A$$

for any $A = (\alpha_{ij}) \in Mat_{n\times m}(K)$. So $\tilde{0} \in Mat_{n\times m}(K)$ and $\tilde{0}$ is a right identity of $Mat_{n\times m}(K)$.

(iv) For any $A = (\alpha_{ij}) \in Mat_{n\times m}(K)$, let $B = (-\alpha_{ij}) \in Mat_{n\times m}(K)$. Then $A + B = (\alpha_{ij}) + (-\alpha_{ij}) = (\alpha_{ij} + (-\alpha_{ij})) = \tilde{0}$. Hence every element $A = (\alpha_{ij})$ in $Mat_{n\times m}(K)$ has an inverse $(-\alpha_{ij})$ in $Mat_{n\times m}(K)$.

(v) For all $A = (\alpha_{ij})$, $B = (\beta_{ij}) \in Mat_{n\times m}(K)$, we have

$$A + B = (\alpha_{ij}) + (\beta_{ij}) = (\alpha_{ij} + \beta_{ij}) = (\beta_{ij} + \alpha_{ij}) = (\beta_{ij}) + (\alpha_{ij}) = B + A$$

and addition on $Mat_{n\times m}(K)$ is commutative.

This proves that $Mat_{n\times m}(K)$ is an abelian group under addition. Now the properties of multiplication by scalars. For any $\alpha,\beta \in K$ and $A = (\alpha_{ij})$, $B = (\beta_{ij}) \in Mat_{n\times m}(K)$, we have

$$\begin{aligned}
(1)\qquad \alpha(A+B) &= \alpha((\alpha_{ij})+(\beta_{ij})) = \alpha(\alpha_{ij}+\beta_{ij}) = (\alpha(\alpha_{ij}+\beta_{ij})) \\
&= (\alpha\alpha_{ij}+\alpha\beta_{ij}) = (\alpha\alpha_{ij})+(\alpha\beta_{ij}) \\
&= \alpha(\alpha_{ij})+\alpha(\beta_{ij}) = \alpha A+\alpha B,
\end{aligned}$$

$$\begin{aligned}
(2)\qquad (\alpha+\beta)A &= (\alpha+\beta)(\alpha_{ij}) = ((\alpha+\beta)\alpha_{ij}) = (\alpha\alpha_{ij}+\beta\alpha_{ij}) \\
&= (\alpha\alpha_{ij})+(\beta\alpha_{ij}) = \alpha(\alpha_{ij})+\beta(\alpha_{ij}) = \alpha A+\beta A,
\end{aligned}$$

$$\begin{aligned}
(3)\qquad (\alpha\beta)A &= (\alpha\beta)(\alpha_{ij}) = ((\alpha\beta)\alpha_{ij}) = (\alpha(\beta\alpha_{ij})) = \alpha(\beta\alpha_{ij}) \\
&= \alpha(\beta(\alpha_{ij})) = \alpha(\beta A),
\end{aligned}$$

$$(4)\qquad 1A = 1(\alpha_{ij}) = (1\alpha_{ij}) = (\alpha_{ij}) = A.$$

Thus $Mat_{n\times m}(K)$ is a vector space over K. □

A convenient K-basis of $Mat_{n\times m}(K)$ is described in the next lemma.

43.5 Lemma: *Let E_{ij} be the matrix in $Mat_{n\times m}(K)$ all of whose entries are 0, except for the single entry in the i-th row, j-th column, which entry is the identity element of K. Then the nm matrices E_{ij} (where $i = 1,2,\ldots,n$; $j = 1,2,\ldots,m$) form a K-basis of $Mat_{n\times m}(K)$. In particular, $dim_K(Mat_{n\times m}(K))$ is equal to nm.*

Proof: The matrices E_{ij} span $Mat_{n\times m}(K)$ over K because any $A = (\alpha_{ij})$ in $Mat_{n\times m}(K)$ can be written as a K-linear combination

$$A = (\alpha_{ij}) = \sum_{i,j} \alpha_{ij}E_{ij}$$

of them. Moreover, matrices E_{ij} are linearly independent over K, for if α_{ij} are scalars such that

$$\sum_{i,j} \alpha_{ij}E_{ij} = \tilde{0},$$

then

$$(\alpha_{ij}) = \tilde{0}$$

and $\alpha_{ij} = 0$ for all i,j. Therefore $\{E_{ij}: i = 1,2,\ldots,n; j = 1,2,\ldots,m\}$ is a K-basis of $Mat_{n\times m}(K)$. In particular, $dim_K(Mat_{n\times m}(K)) = nm$. □

We relate $L(V,W)$ to $Mat_{n\times m}(K)$. This relation is implicit in (*). We state this relation as a definition and prove that $L_K(V,W)$ and $Mat_{n\times m}(K)$ are isomorphic K-vector spaces.

43.6 Definition: Let V be an n-dimensional and W be an m-dimensional vector space over a field K, where $n,m \in \mathbb{N}$. Let $B = \{v_1,v_2,\ldots,v_n\}$ be a K-basis of V and $B' = \{w_1,w_2,\ldots,w_m\}$ be a K-basis of W.
Let T be a K-linear transformation in $L_K(V,W)$ and let

$$v_iT = \sum_{j=1}^{m} \alpha_{ij}w_j \qquad (i = 1,2,\ldots,n), \qquad (*)$$

where $\alpha_{ij} \in K$.
The $n \times m$ matrix (α_{ij}) over K will be called the *matrix associated with T (relative to the bases B and B')*, and will be written $M_B^B(T)$.

In the following discussion, the bases will be fixed and we simply write $M(T)$ instead of $M_B^B(T)$. The role of the bases will be discussed at the end of this paragraph.

43.7 Theorem: *Let V be an n-dimensional and W be an m-dimensional vector space over a field K, where $n,m \in \mathbb{N}$. Then, for any $T,S \in L_K(V,W)$ and $\alpha \in K$, we have*

$$M(T + S) = M(T) + M(S) \qquad \textit{and} \qquad M(\alpha T) = \alpha M(T)$$

(all associated matrices are taken relative to the same pair of K-bases).
In other words, $M: L_K(V,W) \to Mat_{n\times m}(K)$ *is a K-linear transformation.*

Proof: Let $B = \{v_1,v_2,\ldots,v_n\}$ be the K-basis of V and $B' = \{w_1,w_2,\ldots,w_m\}$ be the K-basis of W relative to which the associated matrices are taken, so that $M(T) = (\alpha_{ij})$ and $M(S) = (\beta_{ij})$, where

$$v_i T = \sum_{j=1}^{m} \alpha_{ij} w_j \qquad \text{and} \qquad v_i S = \sum_{j=1}^{m} \beta_{ij} w_j.$$

Then $v_i(T + S) = v_i T + v_i S = \sum_{j=1}^{m} \alpha_{ij} w_j + \sum_{j=1}^{m} \beta_{ij} w_j = \sum_{j=1}^{m} (\alpha_{ij} + \beta_{ij}) w_j$

and therefore $M(T + S) = (\alpha_{ij} + \beta_{ij}) = (\alpha_{ij}) + (\beta_{ij}) = M(T) + M(S)$. Also

$$v_i(\alpha T) = (\alpha v_i)T = \alpha(v_i T) = \alpha \sum_{j=1}^{m} \alpha_{ij} w_j = \sum_{j=1}^{m} \alpha\, \alpha_{ij} w_j$$

and therefore $M(\alpha T) = (\alpha\, \alpha_{ij}) = \alpha(\alpha_{ij}) = \alpha M(T)$. □

43.8 Theorem: *Let V be an n-dimensional and W be an m-dimensional vector space over a field K, where $n,m \in \mathbb{N}$. Then $L_K(V,W)$ is isomorphic to $Mat_{n\times m}(K)$ (as K-vector spaces). In particular, $dim_K L_K(V,W) = nm$.*

Proof: We know that $M: L_K(V,W) \to Mat_{n\times m}(K)$ (in the notation of Theorem 43.7) is a vector space homomorphism. We will prove that M is in fact an isomorphism.

We prove that M is one-to-one and onto. Let (γ_{ij}) be any matrix in $Mat_{n\times m}(K)$. We want to find a T in $L_K(V,W)$ such that $M(T) = (\gamma_{ij})$. Such a K-linear transformation T should satisfy

$$v_i T = \sum_{j=1}^{m} \gamma_{ij} w_j$$

and
$$\Big(\sum_{i=1}^{n} \alpha_i v_i\Big)T = \sum_{i=1}^{n} \alpha_i (v_i T) = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \gamma_{ij} w_j = \sum_{j=1}^{m} \Big(\sum_{i=1}^{n} \alpha_i \gamma_{ij}\Big) w_j$$

for any vector $v = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$ in V. Thus there is at most one T in $L_K(V,W)$ with $M(T) = (\gamma_{ij})$. Hence M is one-to-one.

With the hindsight gained from the chain of equations above, given any (γ_{ij}) in $Mat_{n\times m}(K)$, we *define* a function $T: V \to W$ by

$$\Big(\sum_{i=1}^{n} \alpha_i v_i\Big)T = \sum_{j=1}^{m} \Big(\sum_{i=1}^{n} \alpha_i \gamma_{ij}\Big) w_j.$$

Then, for any $\alpha,\beta \in K$ and $v = \sum_{i=1}^{n} \alpha_i v_i$, $v' = \sum_{i=1}^{n} \beta_i v_i \in V$, we have

$$(\alpha v + \beta v')T = \Big(\alpha \sum_{i=1}^{n} \alpha_i v_i + \beta \sum_{i=1}^{n} \beta_i v_i\Big)T = \Big(\sum_{i=1}^{n} (\alpha\, \alpha_i + \beta\, \beta_i) v_i\Big)T$$

$$= \sum_{j=1}^{m} \Big(\sum_{i=1}^{n} (\alpha\,\alpha_i + \beta\,\beta_i)\gamma_{ij} \Big) w_j = \sum_{j=1}^{m} \Big(\alpha \sum_{i=1}^{n} \alpha_i \gamma_{ij} + \beta \sum_{i=1}^{n} \beta_i \gamma_{ij} \Big) w_j$$

$$= \alpha \sum_{j=1}^{m} \Big(\sum_{i=1}^{n} \alpha_i \gamma_{ij} \Big) w_j + \beta \sum_{j=1}^{m} \Big(\sum_{i=1}^{n} \beta_i \gamma_{ij} \Big) w_j = \alpha(vT) + \beta(v'T)$$

and T is K-linear. Thus $T \in L_K(V,W)$. When we put $\alpha_{i_0} = 1$ and $\alpha_i = 0$ for $i \neq i_0$, we obtain

$$v_{i_0} T = \sum_{j=1}^{m} \gamma_{i_0 j} w_j, \qquad (i_0 = 1,2,\ldots,n)$$

so $M(T) = (\gamma_{ij})$. Thus every $(\gamma_{ij}) \in Mat_{n\times m}(K)$ is the image, under M, of at least one $T \in L_K(V,W)$ and so M is onto. Consequently M is a vector space isomorphism: $L_K(V,W) \cong Mat_{n\times m}(K)$. From Theorem 42.18 and Lemma 43.5, we get $dim_K L_K(V,W) = dim_K Mat_{n\times m}(K) = nm$. □

Now let U be a vector space over the field K, with $dim_K U = k \in \mathbb{N}$, and let $B'' = \{u_1, u_2, \ldots, u_k\}$ be a basis of U over K. If $T: V \to W$ and $S: W \to U$ are K-linear transformations, whose associated matrices [relative to the K-bases $B = \{v_1, v_2, \ldots, v_n\}$, $B' = \{w_1, w_2, \ldots, w_m\}$ of V and W, and relative to the K-bases B', B'' of W and U] are $(\alpha_{ij}) \in Mat_{n\times m}(K)$ and $(\beta_{jl}) \in Mat_{m\times k}(K)$, so that

$$v_i T = \sum_{j=1}^{m} \alpha_{ij} w_j, \qquad w_j S = \sum_{l=1}^{k} \beta_{jl} u_l,$$

then $TS: V \to U$ is a K-linear transformation (Theorem 41.7) and

$$v_i(TS) = (v_i T)S = \Big(\sum_{j=1}^{m} \alpha_{ij} w_j \Big) S = \sum_{j=1}^{m} \alpha_{ij} (w_j S)$$

$$= \sum_{j=1}^{m} \alpha_{ij} \sum_{l=1}^{k} \beta_{jl} u_l = \sum_{l=1}^{k} \Big(\sum_{j=1}^{m} \alpha_{ij} \beta_{jl} \Big) u_l$$

so that the matrix associated with TS [relative to the K-bases B, B''] is the $n \times k$ matrix whose i-th row, l-th column entry is $\sum_{j=1}^{m} \alpha_{ij} \beta_{jl}$. This leads us to the following definition.

43.9 Definition: Let $A = (\alpha_{ij})$ be an $n \times m$ matrix and let $B = (\beta_{ij})$ be an $m \times k$ matrix, with entries from a field K. Then the *product of A and B*, denoted by AB, is the $n \times k$ matrix (γ_{ij}) over K, where $\gamma_{ij} = \sum_{j=1}^{m} \alpha_{ij}\beta_{jl}$. Stated otherwise

$$(\alpha_{ij})(\beta_{ij}) := (\sum_{j=1}^{m} \alpha_{ij}\beta_{jl})$$

Before studying the properties of this matrix multiplication, we summarize the discussion preceding Definition 43.9. Although matrix multiplication is defined in such a way as to make it true, the following theorem is by no means obvious (cf. Remark 43.18).

43.10 Theorem: *Let V,W,U be vector spaces over a field K, of nonzero finite dimensions n,m,k, respectively. Let B, B', B'' be fixed K-bases of V,W,U, respectively. If, relative to these bases, $T \in L_K(V,W)$ has the associated matrix A, and $S \in L_K(W,U)$ has the associated matrix C, then $TS \in L_K(V,U)$ has the associated matrix AC. Equivalently,*

$$M(TS) = M(T)M(S). \qquad \square$$

The product of an $n \times m$ matrix by an $m \times k$ matrix is an $n \times k$ matrix. Notice that the number of columns in A has to be equal to the number of rows in B in order AB to make sense. The product of an $n \times m$ matrix by an $m' \times k$ matrix is *not* defined unless $m = m'$.

Matrix multiplication is associative whenever it is possible. That is to say, $(AB)C = A(BC)$ for any matrices A,B,C with entries from a field K, provided the sizes of A,B,C are such that the products AB and BC are defined (then $(AB)C$ and $A(BC)$ are defined, too). More precisely, if A is an $n \times m$ matrix, B is an $m \times k$ matrix and C is an $k \times s$ matrix, then the two $n \times s$ matrices $(AB)C$, $A(BC)$ are equal. To prove this, let us put $A = (\alpha_{ij})$, $B = (\beta_{jl})$, $C = (\gamma_{lr})$, where $i = 1,2, \ldots,n$; $j = 1,2, \ldots,m$; $r = 1,2, \ldots,s$. Then

$$AB = E = (\epsilon_{il}), \qquad \text{where } \epsilon_{il} = \sum_{j=1}^{m} \alpha_{ij}\beta_{jl}$$

$$BC = F = (\phi_{jr}), \qquad \text{where } \phi_{jr} = \sum_{l=1}^{k} \beta_{jl}\gamma_{lr}$$

and from $(AB)C = EC = \left(\sum_{l=1}^{k} \varepsilon_{il}\gamma_{lr}\right)$

$$= \left(\sum_{l=1}^{k}\left(\sum_{j=1}^{m} \alpha_{ij}\beta_{jl}\right)\gamma_{lr}\right) = \left(\sum_{l=1}^{k}\sum_{j=1}^{m}(\alpha_{ij}\beta_{jl})\gamma_{lr}\right),$$

$$A(BC) = AF = \left(\sum_{j=1}^{m} \alpha_{ij}\phi_{jr}\right) = \left(\sum_{j=1}^{m} \alpha_{ij}\left(\sum_{l=1}^{k} \beta_{jl}\gamma_{lr}\right)\right)$$

$$= \left(\sum_{j=1}^{m}\sum_{l=1}^{k} \alpha_{ij}(\beta_{jl}\gamma_{lr})\right) = \left(\sum_{l=1}^{k}\sum_{j=1}^{m} \alpha_{ij}(\beta_{jl}\gamma_{lr})\right)$$

$$= \left(\sum_{l=1}^{k}\sum_{j=1}^{m}(\alpha_{ij}\beta_{jl})\gamma_{lr}\right),$$

we conclude that $(AB)C = A(BC)$.

However, there is no hope for commutativity. For one thing, the product BA need not be defined even if the product AB happens to be defined. For instance, if A is a 2×3 and B is a 3×4 matrix, then AB is a 2×4 matrix, but BA is not even defined, let alone is equal to AB. But also in cases where both products AB and BA are defined, they will, generally speaking, have diferent sizes, so they will fail to be equal on dimension grounds. For instance, if A is a 2×3 matrix and B is a 3×2 matrix, then AB is a 2×2 matrix and BA is a 3×3 matrix and $AB \neq BA$, since a 2×2 matrix cannot be equal to a 3×3 matrix. Even if both AB and BA are defined and have the same size (this occurs only in case A and B are square matrices with the same number of rows), it usually happens that $AB \neq BA$. For example $\begin{pmatrix}0&0\\1&0\end{pmatrix}\begin{pmatrix}1&0\\0&0\end{pmatrix} \neq \begin{pmatrix}1&0\\0&0\end{pmatrix}\begin{pmatrix}0&0\\1&0\end{pmatrix}$.

Let I_m be the square matrix over K with m rows, whose entries are all equal to $0 \in K$, except for those on the main diagonal, which are all equal to $1 \in K$ (the main diagonal in any $n \times m$ matrix consists of the places where the i-th row and the i-th column intersect, $(i = 1,2, \ldots, \min\{n,m\})$). It is easily verified that $AI_m = A$ for any $A \in Mat_{n\times m}(K)$. Likewise $I_nA = A$ for any $A \in Mat_{n\times m}(K)$.

Let $0_{m\times k}$ be the $m \times k$ matrix over K all of whose entries are $0 \in K$. One checks easily that $A0_{m\times k} = 0_{n\times k}$ and $0_{k\times n}A = 0_{k\times m}$ for any $A \in Mat_{n\times m}(K)$.

Multiplication of matrices is distributive over addition. Indeed, for all $A = (\alpha_{ij}) \in Mat_{n\times m}(K)$ and $B = (\beta_{jl}), C = (\gamma_{jl}) \in Mat_{m\times k}(K)$, we have

$$A(B + C) = (\alpha_{ij})(\beta_{jl} + \gamma_{jl}) = \Big(\sum_{j=1}^{m} \alpha_{ij}(\beta_{jl} + \gamma_{jl})\Big) = \Big(\sum_{j=1}^{m} (\alpha_{ij}\beta_{jl} + \alpha_{ij}\gamma_{jl})\Big)$$

$$= \Big(\sum_{j=1}^{m} \alpha_{ij}\beta_{jl} + \sum_{j=1}^{m} \alpha_{ij}\gamma_{jl}\Big) = \Big(\sum_{j=1}^{m} \alpha_{ij}\beta_{jl}\Big) + \Big(\sum_{j=1}^{m} \alpha_{ij}\gamma_{jl}\Big) = AB + AC.$$

In like manner, one proves $(B + C)A = BA + CA$ for all $A \in Mat_{n\times k}(K)$ and $B, C \in Mat_{m\times n}(K)$.

One checks easily that, for all $\alpha \in K$, $A \in Mat_{n\times m}(K)$, $B \in Mat_{m\times k}(K)$

$$(\alpha A)B = \alpha(AB) = A(\alpha B). \tag{e}$$

On writing $A = (\alpha_{ij})$, $B = (\beta_{jl})$, we get indeed

$$(\alpha A)B = (\alpha\alpha_{ij})(\beta_{jl}) = \Big(\sum_{j=1}^{m} (\alpha\alpha_{ij})\beta_{jl}\Big) = \Big(\sum_{j=1}^{m} \alpha(\alpha_{ij}\beta_{jl})\Big)$$

$$= \Big(\alpha\sum_{j=1}^{m} \alpha_{ij}\beta_{jl}\Big) = \alpha\Big(\sum_{j=1}^{m} \alpha_{ij}\beta_{jl}\Big) = \alpha(AB)$$

and

$$A(\alpha B) = (\alpha_{ij})(\alpha\beta_{jl}) = \Big(\sum_{j=1}^{m} \alpha_{ij}(\alpha\beta_{jl})\Big) = \Big(\sum_{j=1}^{m} \alpha(\alpha_{ij}\beta_{jl})\Big)$$

$$= \Big(\alpha\sum_{j=1}^{m} \alpha_{ij}\beta_{jl}\Big) = \alpha\Big(\sum_{j=1}^{m} \alpha_{ij}\beta_{jl}\Big) = \alpha(AB).$$

Let us now consider the set $Mat_n(K)$ of square matrices over a field K. From Theorem 43.3, we know that $Mat_n(K)$ is an abelian group under addition. The product of any two $n \times n$ matrices is an $n \times n$ matrix. Since the matrix multiplication is associative and distributive over addition, $Mat_n(K)$ is a ring. We also know that $AI_n = I_nA = A$ for any $n \times n$ matrix A. Thus we proved

43.11 Theorem: *Let K be a field and $n \in \mathbb{N}$. Then, under matrix addition and matrix multiplication, $Mat_n(K)$ is a ring with identity I_n.* □

The counterpart of Theorem 43.11 for linear transformations is also valid.

43.12 Theorem: *Let V be a vector space over a field K and let $L_K(V,V)$ be the set of all K-linear mappings from V into V. Then, under the pointwise addition and composition of K-linear transformations, $L_K(V,V)$ is a ring with identity. The identity mapping $\imath: V \to V$ is the identity element of this ring $L_K(V,V)$.* [Notice that there is no hypothesis about $dim_K V$.]

Proof: We must check the ring axioms. From Theorem 43.1, we know that $L_K(V,V)$ is an abelian group under addition. Also, (1) $L_K(V,V)$ is closed under the composition of mappings (Theorem 41.7), and (2) composition of mappings (whether K-linear or not) is associative (Theorem 3.10), and (D) composition is distributive over addition: when T,S,R are arbitrary elements of $L_K(V,V)$, then

$$v(T(S+R)) = (vT)(S+R) = ((vT)S) + ((vT)R) = (v(TS)) + (v(TR)) = v(TS+TR)$$

and $$v((S+R)T) = (v(S+R))T = (vS+vR)T = (vS)T + (vR)T = v(ST) + v(RT)$$
$$= v(ST+RT)$$

for all $v \in V$, hence $T(S+R) = TS + TR$ and $(S+R)T = ST + RT$. So $L_K(V,V)$ is a ring. Finally, the identity mapping $\imath$ is clearly a K-linear transformation, so $\imath \in L_K(V,V)$ and as $T\imath = T = \imath T$ for all $T \in L_K(V,V)$, we conclude that $L_K(V,V)$ is a ring with identity $\imath$. □

43.13 Theorem: *Let V be a vector space over a field K with $dim_K V = n$, where $n \in \mathbb{N}$. Then $L_K(V,V) \cong Mat_n(K)$ (ring isomorphism).*

Proof: We fix a K-basis of V and use the mapping $M: L_K(V,V) \to Mat_n(K)$ of Theorem 43.7, so that $M(T)$ is the associated matrix of the K-linear transformation $T \in L_K(V,V)$. By Theorem 43.8, M is an isomorphism of abelian groups (in fact of K-vector spaces, but we do not need this now) and by Theorem 43.10, M preserves multiplication as well. Hence M is a ring isomorphism. □

Let us recall that a unit in a ring with identity is an element of that ring possessing a (unique) right inverse which is also a left inverse. What are the units of $L_K(V,V)$? The units in $L_K(V,V)$ are, by definition, those K-linear transformations T with the inverse T^{-1} in $L_K(V,V)$. The inverse of T in $L_K(V,V)$, whenever it exists, is in $L_K(V,V)$ by Lemma 41.10(2). Thus the units of $L_K(V,V)$ are the K-linear transformations in $L_K(V,V)$ which are one-to-one and onto: the units of $L_K(V,V)$ are the vector space isomorphisms from V onto V. The set of all isomorphisms from V onto V will be denoted by $GL(V)$. This is a group under the composition of mappings, called the *general linear group of* V. Thus $L_K(V,V)^\times = GL(V)$.

The units in $Mat_n(K)$ are the invertible matrices, that is to say, matrices A in $Mat_n(K)$ for which an $A^{-1} \in Mat_n(K)$ exists such that $AA^{-1} = I_n = A^{-1}A$. These are the matrices associated with isomorphisms from V onto V. In the next paragraph, we will give a necessary and sufficient condition for a matrix to be invertible (Theorem 44.20). The set of all invertible matrices in $Mat_n(K)$ will be denoted by $GL(n,K)$. This is a group under the multiplication of matrices, called the *general linear group of degree* n *over* K. Thus $Mat_n(K)^\times = GL(n,K)$. When V is an n-dimensional vector space over K, the group $GL(V)$ is isomorphic to the group $GL(n,K)$.

We return to the more general case of $L_K(V,W)$ and $Mat_{n\times m}(K)$. Suppose again that V and W are K-vector spaces of K-dimensions n and m, respectively, where $n,m \in \mathbb{N}$. Let $B = \{v_1,v_2,\ldots,v_n\}$ and $B^* = \{v_1^*,v_2^*,\ldots,v_n^*\}$ be K-bases of V and let $B' = \{w_1,w_2,\ldots,w_m\}$ and $B'^* = \{w_1^*,w_2^*,\ldots,w_m^*\}$ be K-bases of W. With each K-linear transformation $T: V \to W$, there is associated a matrix $M_B^{B'}(T)$ relative to the bases B and B', and a matrix $M_{B^*}^{B'^*}(T)$ relative to the bases B^* and B'^*. We want to study the relationship between $M_B^{B'}(T)$ and $M_{B^*}^{B'^*}(T)$.

We recall that $M_B^{B'}(T) = (\alpha_{ij})$ and $M_{B^*}^{B'^*}(T) = (\beta_{ij})$, where

$$v_iT = \sum_{j=1}^{m} \alpha_{ij}w_j \quad \text{and} \quad v_i^*T = \sum_{j=1}^{m} \beta_{ij}w_j^*. \qquad (i= 1,2,\ldots,n)$$

We introduce transition matrices which describe the change of bases. Writing

$$v_i = \sum_{k=1}^{n} \gamma_{ik} v_k^* \qquad (i = 1,2, \ldots, n)$$

we obtain a matrix (γ_{ik}) in $Mat_n(K)$, called the *transition matrix from the K-basis B to the K-basis B* of V*. Of course, $(\gamma_{ik}) = M_{B^*}^{B}(\iota)$, where ι is the identity mapping on V. We have the schema

$$\begin{array}{ccccc} & \overset{\iota}{\rightarrow} & & \overset{\iota}{\rightarrow} & \\ V & & V & & V \\ B^* & & B & & B^* \\ & M_B^{B^*}(\iota) & & M_{B^*}^{B}(\iota) & \end{array} \qquad \begin{array}{l} \text{mappings} \\ \text{vector spaces} \\ \text{bases} \\ \text{matrices} \end{array}$$

Now the composition $\iota\iota$ is the identity mapping ι. Relative to the bases B^* and B^*, the matrix associated with ι is the identity matrix I_n. This matrix is also equal to $M_B^{B^*}(\iota)M_{B^*}^{B}(\iota)$ by Theorem 43.10: $M_B^{B^*}(\iota)M_{B^*}^{B}(\iota) = I_n$. Thus the transition matrix from B to B^* is the inverse of the transition matrix from B^* to B.

$$\begin{array}{ccc} & \overset{\iota}{\rightarrow} & \\ V & & V \\ B & & B^* \\ & M_{B^*}^{B}(\iota) & \end{array} \qquad \begin{array}{ccc} & \overset{\iota}{\rightarrow} & \\ V & & V \\ B^* & & B \\ & M_B^{B^*}(\iota) = (M_{B^*}^{B}(\iota))^{-1} & \end{array}$$

43.14 Theorem: *With the foregoing notation, let P be the transition matrix from B to B*, and let Q be the transition matrix from B′ to B′*. If $T: V \to W$ is any K-linear mapping, then the matrices $M_B^B(T)$ and $M_{B'^*}^{B^*}(T)$ are connected by $M_{B'^*}^{B^*}(T) = P^{-1}M_B^B(T)Q$.*

Proof: We have the following schema

$$\begin{array}{ccc} & \overset{T}{\rightarrow} & \\ V & & W \\ B & & B' \\ & M_B^B(T) & \end{array}$$

The K-linear transformation T can be desribed also as follows.

$$\begin{array}{ccccccc} & \overset{\iota_V}{\rightarrow} & & \overset{T}{\rightarrow} & & \overset{\iota_W}{\rightarrow} & \\ V & & V & & W & & W \\ B^* & & B & & B' & & B'^* \\ & M_B^{B^*}(\iota_V) & & M_B^B(T) & & M_{B'^*}^{B'}(\iota_W) & \end{array}$$

Now $M_B^{B^*}(\iota_V) = [M_{B^*}^B(\iota_V)]^{-1} = P^{-1}$ and $M_{B'^*}^{B'}(\iota_W) = Q$. By Theorem 43.10, the matrix associated with T relative to the bases B^* and B'^* is $P^{-1}M_{B'}^B(T)Q$. Hence $M_{B'^*}^{B^*}(T) = P^{-1}M_{B'}^B(T)Q$.

$$\begin{array}{ccccccc} & & \iota_V & & T & & \iota_W \\ V & \to & V & \to & W & \to & W \\ B^* & & B & & B' & & B'^* \\ & & P^{-1} & & M_{B'}^B(T) & & Q \end{array}$$

□

43.15 Theorem: *Let V be an n-dimensional vector space over a field K, B and B^* be K-bases of V, and let P be the transition matrix from B to B^*. Suppose T is any K-linear mapping from V into V. If $M(T)$ is the matrix associated with T relative to the bases B and B, and if $M^*(T)$ is the matrix associated with T relative to the bases B^* and B^*, then*

$$M^*(T) = P^{-1}M(T)P.$$

Proof: This is a special case of Theorem 43.14. Using the diagram

$$\begin{array}{ccccccc} & & \iota & & T & & \iota \\ V & \to & V & \to & V & \to & V \\ B^* & & B & & B & & B^* \\ & & P^{-1} & & M(T) & & P \end{array}$$

the proof follows immediately from Theorem 43.10. □

43.16 Definition: Let K be a field and let $A = (\alpha_{ij})$ be an $n \times m$ matrix with entries from K. Then the $m \times n$ matrix whose j-th row, i-th column entry is equal to α_{ij} is called the *transpose of* A, and is written A^t.

Hence A^t is obtained from A by changing rows to columns and columns to rows. For instance, the transpose of $\begin{pmatrix} 0 & 1 & 2 \\ 1 & 3 & 5 \end{pmatrix}$ is $\begin{pmatrix} 0 & 1 \\ 1 & 3 \\ 2 & 5 \end{pmatrix}$. It follows from the definition that $(A^t)^t = A$ for any matrix A.

43.17 Lemma: *Let K be a field and let $A,B \in Mat_{n\times m}(K)$, $C \in Mat_{m\times k}(K)$, $\alpha \in K$. Then*

$$(A+B)^t = A^t + B^t, \qquad (\alpha A)^t = \alpha(A^t), \qquad (AC)^t = C^tA^t.$$

Proof: Let $A = (\alpha_{ij})$, $B = (\beta_{ij})$ and $C = (\gamma_{jl})$, where $i = 1,2,\ldots,n$; $j = 1,2,\ldots,m$ and $l = 1,2,\ldots,k$. Then $A^t = (\alpha_{ji})$, $B^t = (\beta_{ji})$, $C^t = (\gamma_{lj})$ and $\alpha A = (\alpha\,\alpha_{ij})$. So

$$\begin{aligned}(A+B)^t &= ((\alpha_{ij}) + (\beta_{ij}))^t = (\alpha_{ij}+\beta_{ij})^t \\ &= \text{matrix whose } j\text{-th row, } i\text{-th column entry is } \alpha_{ij}+\beta_{ij} \\ &= (\text{matrix whose } j\text{-th row, } i\text{-th column entry is } \alpha_{ij}) \\ &\quad + (\text{matrix whose } j\text{-th row, } i\text{-th column entry is } \beta_{ij}) \\ &= (\alpha_{ij})^t + (\beta_{ij})^t = A^t + B^t,\end{aligned}$$

$$\begin{aligned}(\alpha A)^t = (\alpha\,\alpha_{ij})^t &= \text{matrix whose } j\text{-th row, } i\text{-th column entry is } \alpha\,\alpha_{ij} \\ &= \alpha(\text{matrix whose } j\text{-th row, } i\text{-th column entry is } \alpha_{ij}) \\ &= \alpha(\alpha_{ij})^t = \alpha A^t,\end{aligned}$$

$$\begin{aligned}(AC)^t &= \text{matrix whose } l\text{-th row, } i\text{-th column entry is the } i\text{-th row, } l\text{-th column entry in } AC \\ &= \text{matrix whose } l\text{-th row, } i\text{-th column entry is } \sum_{j=1}^{m}\alpha_{ij}\,\gamma_{jl} \\ &= \text{matrix whose } l\text{-th row, } i\text{-th column entry is } \sum_{j=1}^{m}\gamma_{jl}\,\alpha_{ij} \\ &= (\text{matrix whose } l\text{-th row, } j\text{-th column entry is } \gamma_{jl}) \text{times} \\ &\quad (\text{matrix whose } j\text{-th row, } i\text{-th column entry is } \alpha_{ij}) \\ &= (\gamma_{ij})^t(\alpha_{ij})^t = C^tA^t.\end{aligned}$$

□

43.18 Remark: The results in this paragraph are very natural. All operations discussed here are natural, and the vector spaces and rings of this paragraph arise naturally. Another natural item is the isomorphism in Theorem 43.10.

There is, however, a subtle point here. Theorem 43.10 is true only because we write the functions on the right of the elements on which

they act! If we had written them on the left, Theorem 43.10 would read: $M(TS) = M(S)M(T)$. Of course this is not as good as $M(TS) = M(T)M(S)$. For this reason, people who write functions on the left define the associated matrices differently. If $T \in L_K(V,W)$ and $v_iT = \sum_{j=1}^{m} \alpha_{ij} w_j$ as in Definition 43.6, they define the matrix associated with T (relative to the fixed K-bases $\{v_1, v_2, \ldots, v_n\}$ of V and $\{w_1, w_2, \ldots, w_m\}$ of W) to be $(\alpha_{ij})^t$. Thus their $M(T)$ is our $M(T)^t$, and

their $M(TS)$ = their M(first S, then T) = our M(first S, then T)^t = our $M(ST)^t$ = our $(M(S)M(T))^t$ = our $M(T)^tM(S)^t$ = their $M(T)M(S)$

so that Theorem 43.10 is true in their notation, too. In some books, the forming of the transpose is included in the notation for the associated matrix. More clearly, some people write

$$v_iT = \sum_{j=1}^{m} \alpha_{ji} w_j$$

and define the matrix associated with T to be (α_{ji}). Then $M(TS) = M(T)M(S)$ as before, but the equations above are not very sensible, for α_{ji} depends primarily on i and secondarily on j, so the indices occupy wrong places. In our notation, there is no need for the artificial transpositions, nor do we write the indices in the wrong order.

Exercises

1. Compute $A + B$ and AB when

$$A = \begin{pmatrix} 0 & 1 & 3 & 4 \\ -2 & 4 & -2 & 1 \\ 1 & 0 & 3 & 0 \\ 6 & 1 & -1 & 2 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 7 & 1 & -3 \\ 1 & 0 & 3 & -4 \\ 2 & 0 & 2 & 1 \\ 0 & -2 & 1 & 0 \end{pmatrix} \text{ in } Mat_4(\mathbb{R}),$$

and when

$$A = \begin{pmatrix} \bar{5} & \bar{3} & \bar{2} & \bar{4} \\ \bar{3} & \bar{4} & \bar{2} & \bar{1} \\ \bar{1} & \bar{2} & \bar{3} & \bar{1} \\ \bar{0} & \bar{4} & \bar{6} & \bar{1} \end{pmatrix}, \qquad B = \begin{pmatrix} \bar{1} & \bar{2} & \bar{5} & \bar{1} \\ \bar{0} & \bar{3} & \bar{3} & \bar{6} \\ \bar{2} & \bar{6} & \bar{1} & \bar{4} \\ \bar{0} & \bar{4} & \bar{5} & \bar{0} \end{pmatrix} \text{ in } Mat_4(\mathbb{Z}_7).$$

2. Let K be a field and $A,B \in Mat_n(K)$ with $AB = BA$. Prove that $(A + B)^2 = A^2 + 2AB + B^2$ and $(A + B)(A - B) = A^2 - B^2$. Show that these equations need not hold if $AB \neq BA$.

3. Evaluate $A,A^2,A^3, \ldots$, where A is given by

$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ and by $A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$. Generalize to square matrices of n rows.

4. Evaluate $A,A^2,A^3, \ldots$, where A is given by

$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ and by $A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. Generalize to square matrices of n rows.

4. The *trace* of a matrix $A = (\alpha_{ij}) \in Mat_n(K)$ is defined to be the sum of the entries on the main diagonal of A, denoted by $tr(A)$, so that $tr(A) = \alpha_{11} + \alpha_{22} + \ldots + \alpha_{nn}$. Prove that $tr(A) = tr(A^t)$, that $tr(AB) = tr(BA)$ and that $tr(C^{-1}AC) = tr(A)$ for any $A,B \in Mat_n(K)$, $C \in GL(n,K)$.

6. Let V,W be vector spaces over a field K, let $\{v_i : i \in I\}$ be a K-basis of V and let $T,S \in L_K(V,W)$. Prove that $T = S$ if and only if $v_iT = v_iS$ for all $i \in I$.

7. Let V be a vector space over a field K, with $dim_K V = n \in \mathbb{N}$. Prove that $GL(V)$ is isomorphic to $GL(n,K)$.

8. Let $\varphi: \mathbb{R}^3 \to \mathbb{R}^3$ be the $\mathbb{R}$-linear mapping for which $u_1\varphi = (1,0,2)$, $u_2\varphi = (0,1,1)$, $u_3\varphi = (1,0,1)$, where, as usual, $u_1 = (1,0,0)$, $u_2 = (0,1,0)$, $u_3 = (0,0,1)$. We put $v_1 = (-1,1,0)$, $v_2 = (1,2,3)$, $v_3 = (0,1,2)$. Let $B = \{u_1,u_2,u_3\}$, and $B^* = \{v_1,v_2,v_3\}$. Show that B^* is an $\mathbb{R}$-basis of $\mathbb{R}^3$ and find the matrix of the $\mathbb{R}$-linear transformation φ relative to the bases (a) B and B; (b) B and B^*; (c) B^* and B; (d) B^* and B^*.

§44
Determinants

With each (square) matrix over a field K, we associate an element of K, called the determinant of the matrix. In this paragraph, we study the properties of determinants.

Determinants arise in many contexts. For example, if a_1,a_2,b_1,b_2 elements of a field K and if the equations

$$a_1x + a_2y = 0$$
$$b_1x + b_2y = 0$$

hold, then $(a_1b_2 - a_2b_1)x = (a_1b_2 - a_2b_1)y = 0$. In §17, we called $a_1b_2 - a_2b_1$ the determinant of the matrix $\begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}$.

Now let a_3,b_3,c_1,c_2,c_3 be further elements of K. If

$$a_1x + a_2y + a_3z = 0$$
$$b_1x + b_2y + b_3z = 0$$
$$c_1x + c_2y + c_3z = 0,$$

then, multiplying the first equation by $b_2c_3 - b_3c_2$, the second by $a_3c_2 - a_2c_3$, the third by $a_2b_3 - a_3b_2$ and adding them, we get $Dx = 0$, where

$$D = a_1b_2c_3 - a_1b_3c_2 + a_3b_1c_2 - a_2b_1c_3 + a_2b_3c_1 - a_3b_2c_1.$$

One obtains also $Dy = 0$ and $Dz = 0$. Here D is a sum of $6 = 3!$ terms $\mp a_ib_jc_k$, where $\{i,j,k\} = \{1,2,3\}$ and the sign of $a_ib_jc_k$ is + or − according as $\begin{pmatrix} 1 & 2 & 3 \\ i & j & k \end{pmatrix}$ is an even or odd permutation in S_3.

Similarly, when we try to eliminate x,y,z,u from the equations

$$a_1x + a_2y + a_3z + a_4u = 0$$
$$b_1x + b_2y + b_3z + b_4u = 0$$
$$c_1x + c_2y + c_3z + c_4u = 0$$
$$d_1x + d_2y + d_3z + d_4u = 0,$$

we get $D'x = D'y = D'z = D'u = 0$, where D is a sum of $24 = 4!$ terms $\mp a_i b_j c_k d_l$, where $\{i,j,k,l\} = \{1,2,3,4\}$ and the sign of $a_i b_j c_k d_l$ is + or − according as $\begin{pmatrix} 1 & 2 & 3 & 4 \\ i & j & k & l \end{pmatrix}$ is an even or odd permutation in S_4.

This pattern continues. The expressions we get in this way are called determinants. On changing a to α_1, b to α_2, c to α_3, etc., the formal definition reads as follows.

44.1 Definition: Let K be a field and

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} \end{pmatrix} = (\alpha_{ij})$$

be an $n \times n$ square matrix with entries from K. Then the element

$$\sum_{\sigma \in S_n} \varepsilon(\sigma) \alpha_{1,1\sigma} \alpha_{2,2\sigma} \cdots \alpha_{n,n\sigma}$$

of K is called the *determinant of the matrix* A. It will be denoted as

$$det\ A, \qquad \text{or } det(A), \quad \text{or } \begin{vmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} \end{vmatrix}, \text{ or } |\alpha_{ij}|.$$

Hence $det\ A$ is a sum of $n!$ terms These summands are obtained from the product $\alpha_{11}\alpha_{22}\cdots\alpha_{nn}$ of the entries in the main diagonal by permuting the second indices in all the $n!$ ways and attaching a "+" or "−" sign according as the permutation is even or odd. Each summand, aside from its sign, is the product of n entries of the matrix, the entries being from distinct rows and distinct columns. The determinant can also be written

$$\sum_{\sigma \in A_n} \alpha_{1,1\sigma} \alpha_{2,2\sigma} \cdots \alpha_{n,n\sigma} - \sum_{\sigma \in S_n \setminus A_n} \alpha_{1,1\sigma} \alpha_{2,2\sigma} \cdots \alpha_{n,n\sigma}.$$

44.2 Remarks: (1) Determinants are defined for square matrices only. Nonsquare matrices do not have a determinant. Note that the determinant of the 1×1 matrix (α) is equal to $\alpha \in K$.

(2) Definition 44.1 makes sense when K is merely a commutative ring. The theory in this paragraph extends immediately to the case where K is a commutative ring with identity. We will not need this general theory. We observe only: when R is a subring of K, and all entries of $A \in Mat_n(K)$ are in R, then $det\ A$ is in fact an element of R.

Some fundamental properties of determinants are collected in the next lemmas.

44.3 Lemma: *Let K be a field and $A \in Mat_n(K)$. Then $det\ A = det\ A^t$.* (The determinant does not change when rows are changed to columns.)

Proof: Let $A = (\alpha_{ij})$ and $A^t = (\beta_{ij})$, so that $\alpha_{ij} = \beta_{ji}$ for all i,j. Then

$$det\ A = \sum_{\sigma \in S_n} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma}.$$

As σ runs through S_n, so does σ^{-1}. Hence

$$det\ A = \sum_{\sigma \in S_n} \varepsilon(\sigma^{-1})\alpha_{1,1\sigma^{-1}}\alpha_{2,2\sigma^{-1}}\cdots\alpha_{n,n\sigma^{-1}}.$$

Using commutativity of multiplication in K, we reorder the factors in each summand with regard to their second indices and get

$$det\ A = \sum_{\sigma \in S_n} \varepsilon(\sigma^{-1})\alpha_{1\sigma,1}\alpha_{2\sigma,2}\cdots\alpha_{n\sigma,n}.$$

Since $\varepsilon(\sigma^{-1}) = \varepsilon(\sigma)$ for all $\sigma \in S_n$ (in case $n \geqslant 2$; if $n = 1$, there is nothing to prove, for then $A = A^t$),

$$\begin{aligned} det\ A &= \sum_{\sigma \in S_n} \varepsilon(\sigma)\alpha_{1\sigma,1}\alpha_{2\sigma,2}\cdots\alpha_{n\sigma,n} \\ &= \sum_{\sigma \in S_n} \varepsilon(\sigma)\beta_{1,1\sigma}\beta_{2,2\sigma}\cdots\beta_{n,n\sigma} \\ &= det\ A^t. \end{aligned}$$

□

44.4 Lemma: *Let K be a field and $A \in Mat_n(K)$. If each element in a particular row (column) of A is multiplied by $\gamma \in K$, then the determinant of the new matrix thus obtained is equal to $\gamma \cdot \det A$.*

Proof: In view of Lemma 44.3, it suffices to prove the statement about rows only. Let $A = (\alpha_{ij})$. Assume that the elements of the k-th row are multiplied by γ. The new matrix is (β_{ij}), where $\beta_{ij} = \alpha_{ij}$ for $i \neq k$ and $\beta_{kj} = \gamma\alpha_{kj}$. Thus

$$\begin{aligned} |\beta_{ij}| &= \sum_{\sigma \in S_n} \varepsilon(\sigma)\beta_{1,1\sigma}\dots\beta_{k,k\sigma}\dots\beta_{n,n\sigma} \\ &= \sum_{\sigma \in S_n} \varepsilon(\sigma)\alpha_{1,1\sigma}\dots(\gamma\alpha_{k,k\sigma})\dots\alpha_{n,n\sigma} \\ &= \gamma \sum_{\sigma \in S_n} \varepsilon(\sigma)\alpha_{1,1\sigma}\dots\alpha_{k,k\sigma}\dots\alpha_{n,n\sigma} \\ &= \gamma \cdot |\alpha_{ij}|. \end{aligned}$$

□

44.5 Lemma: *Let K be a field and $A \in Mat_n(K)$. Assume that each element α_{kj} in the k-th row of A is a sum $\beta_{kj} + \gamma_{kj}$ (each element α_{kj} in the k-th column of A is a sum $\beta_{ik} + \gamma_{ik}$). Then $\det A$ is a sum of two determinants:*

$$\det A = \det B + \det C,$$

where B resp. C is identical with A, except for the k-th row (column), in which α_{kj} are replaced by β_{kj} resp. γ_{kj} (α_{ik} are replaced by β_{ik} resp. γ_{ik}). Symbolically

$$\begin{vmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \dots & \dots & \dots & \dots \\ \beta_{k1}+\gamma_{k1} & \beta_{k2}+\gamma_{k2} & \dots & \beta_{kn}+\gamma_{kn} \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{vmatrix}$$

$$= \begin{vmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \dots & \dots & \dots & \dots \\ \beta_{k1} & \beta_{k2} & \dots & \beta_{kn} \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{vmatrix} + \begin{vmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \dots & \dots & \dots & \dots \\ \gamma_{k1} & \gamma_{k2} & \dots & \gamma_{kn} \\ \dots & \dots & \dots & \dots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{vmatrix}$$

and

$$\begin{vmatrix} \alpha_{11} & \cdots & \beta_{1k} + \gamma_{1k} & \cdots & \alpha_{1n} \\ \alpha_{21} & \cdots & \beta_{2k} + \gamma_{2k} & \cdots & \alpha_{2n} \\ & & \cdots\cdots\cdots & & \\ \alpha_{n1} & \cdots & \beta_{nk} + \gamma_{nk} & \cdots & \alpha_{nn} \end{vmatrix}$$

$$= \begin{vmatrix} \alpha_{11} & \cdots & \beta_{1k} & \cdots & \alpha_{1n} \\ \alpha_{21} & \cdots & \beta_{2k} & \cdots & \alpha_{2n} \\ & & \cdots\cdots\cdots & & \\ \alpha_{n1} & \cdots & \beta_{nk} & \cdots & \alpha_{nn} \end{vmatrix} + \begin{vmatrix} \alpha_{11} & \cdots & \gamma_{1k} & \cdots & \alpha_{1n} \\ \alpha_{21} & \cdots & \gamma_{2k} & \cdots & \alpha_{2n} \\ & & \cdots\cdots\cdots & & \\ \alpha_{n1} & \cdots & \gamma_{nk} & \cdots & \alpha_{nn} \end{vmatrix}.$$

Proof: The proof is shorter than the wording of the lemma. It will be sufficient to prove the assertion involving rows only, and this follows from summing

$$\begin{aligned} &\varepsilon(\sigma)\alpha_{1,1\sigma}\cdots\alpha_{k,k\sigma}\cdots\alpha_{n,n\sigma} \\ &= \varepsilon(\sigma)\alpha_{1,1\sigma}\cdots(\beta_{k,k\sigma} + \gamma_{k,k\sigma})\cdots\alpha_{n,n\sigma} \\ &= \varepsilon(\sigma)\alpha_{1,1\sigma}\cdots\beta_{k,k\sigma}\cdots\alpha_{n,n\sigma} + \varepsilon(\sigma)\alpha_{1,1\sigma}\cdots\gamma_{k,k\sigma}\cdots\alpha_{n,n\sigma} \end{aligned}$$

over all $\sigma \in S_n$. □

The last two lemmas mean that the determinant of a matrix is a linear function of any one of its rows or columns.

44.6 Lemma: *Let K be a field, $A \in Mat_n(K)$ and $\gamma \in K$. Then det $(\gamma A) = \gamma^n \cdot det\ A$.*

Proof: This follows from n successive applications of Lemma 44.4. Altenatively, observe that, when we put $A = (\alpha_{ij})$, each summand $\varepsilon(\sigma)(\gamma\alpha_{1,1\sigma})(\gamma\alpha_{2,2\sigma})\cdots(\gamma\alpha_{n,n\sigma})$ of *det* (γA) is γ^n times a summand $\varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma}$ of *det* A, and conversely. □

44.7 Lemma: *Let K be a field and $A, B \in Mat_n(K)$. If B is obtained from A by interchanging two rows (columns) of A, then det $B = -$ det A* (the determinant changes sign when two rows (columns) are interchanged.)

Proof: We prove the statement about rows only. Assume $A = (\alpha_{ij})$ and $B = (\beta_{ij})$, and assume that B is obtained from A by interchanging the k-th and m-th rows of A so that $\beta_{ij} = \alpha_{ij}$ for all i,j with $i \neq k$, $i \neq m$ and $\beta_{kj} = \alpha_{mj}$, $\beta_{mj} = \alpha_{kj}$ for all j. Then

$$det\ B = \sum_{\sigma \in S_n} \varepsilon(\sigma)\beta_{1,1\sigma} \cdots \beta_{k,k\sigma} \cdots \beta_{m,m\sigma} \cdots \beta_{n,n\sigma}.$$

As σ ranges over S_n, so does $(km)\sigma$. Hence we have

$$\begin{aligned} det\ B &= \sum_{\sigma \in S_n} \varepsilon((km)\sigma)\beta_{1,1(km)\sigma} \cdots \beta_{k,k(km)\sigma} \cdots \beta_{m,m(km)\sigma} \cdots \beta_{n,n(km)\sigma} \\ &= -\sum_{\sigma \in S_n} \varepsilon(\sigma)\beta_{1,1\sigma} \cdots \beta_{k,m\sigma} \cdots \beta_{m,k\sigma} \cdots \beta_{n,n\sigma} \\ &= -\sum_{\sigma \in S_n} \varepsilon(\sigma)\alpha_{1,1\sigma} \cdots \alpha_{m,m\sigma} \cdots \alpha_{k,k\sigma} \cdots \alpha_{n,n\sigma} \\ &= -\ det\ A. \end{aligned}$$

□

44.8 Lemma: *Let K be a field and $A,B \in Mat_n(K)$, $n \geqslant 2$. If B is obtained from A by a permutation τ of the rows (columns) of A, then $det\ B = \varepsilon(\tau) det\ A$.*

Proof: Let $A = (\alpha_{ij})$ and $B = (\beta_{ij})$. We give a proof of the assertion about rows only. The hypothesis is that $\beta_{ij} = \alpha_{i\tau,j}$ for some τ in S_n. We write τ as a product of transpositions:

$$\tau = \tau_1\tau_2 \cdots \tau_s \qquad (\tau_1,\tau_2,\ldots,\tau_s \text{ are transpositions in } S_n)$$

so that $\varepsilon(\tau) = (-1)^s$ by definition. We introduce matrices

$$A = A_0, A_1, A_2, \ldots, A_{s-1}, A_s = B,$$

where each A_r is obtained from A_{r-1} $(r = 1,2,\ldots,s)$ by interchanging two rows:

$$A_0 = (\alpha_{ij}),\ A_1 = (\alpha_{i\tau_1,j}),\ A_2 = (\alpha_{i\tau_1\tau_2,j}),\ \ldots,\ A_{s-1} = (\alpha_{i\tau_1\tau_2\ldots\tau_{s-1},j}),\ A_s = (\alpha_{i\tau_1\tau_2\ldots\tau_{s-1}\tau_s,j}).$$

Then, using Lemma 44.7 repeatedly,

$$\begin{aligned} det\ B = det\ A_s = -\ det\ A_{s-1} &= (-1)^2 det\ A_{s-2} = (-1)^3 det\ A_{s-3} \\ &= \cdots = (-1)^s det\ A_0 = \varepsilon(\tau) det\ A. \end{aligned}$$

□

44.9 Lemma: *Let K be a field and $A \in Mat_n(K)$, $n \geqslant 2$. If two rows (columns) of A are identical, then $det\ A = 0$.*

Proof: One usually argues as follows. Interchanging the two identical rows (columns), the *det A* does not change. But it becomes $-$ *det A* by Lemma 44.7. Hence *det A* $=$ $-$ *det A*. Thus $2det\ A = 0$. One concludes from this that *det A* $= 0$.

This conclusion is justified when we can divide by 2 in K, that is to say, if the multiplicative inverse of 2 exists in K. Let us recall that 2 is an abbreviation of $1_K + 1_K$, where 1_K is the identity of K. Since any nonzero element of K has an inverse in K, the conclusion is valid when K is a field in which $1_K + 1_K \neq 0$. If, however, $1_K + 1_K = 0$ (as in $\mathbb{Z}_2$), this argument does not work.

We give an argument which works irrespective of whether $1_K + 1_K = 0$ or not. We prove the statement about rows only. Reordering the rows of A by a suitable permutation in S_n, we obtain a matrix B in which the first two rows are identical and $det\ B = \varepsilon(\tau) det\ A$. Since *det B* $= 0$ if and only if *det A* $= 0$, we may assume, without loss of generality, that the first and second rows of A are identical. We prove *det A* $= 0$ under this assumption.

Let $A = (\alpha_{ij})$, with $\alpha_{1j} = \alpha_{2j}$ for all j. If $n = 2$, then $det\ A = \begin{vmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{11} & \alpha_{12} \end{vmatrix} = \alpha_{11}\alpha_{12} - \alpha_{12}\alpha_{11} = 0$. Let us suppose now $n \geqslant 3$. Then

$$det\ A = \sum_{\sigma \in A_n} \alpha_{1,1\sigma}\alpha_{2,2\sigma}\alpha_{3,3\sigma}\cdots\alpha_{n,n\sigma} - \sum_{\sigma \in S_n \setminus A_n} \alpha_{1,1\sigma}\alpha_{2,2\sigma}\alpha_{3,3\sigma}\cdots\alpha_{n,n\sigma}. \quad (i)$$

As σ runs through A_n, the permutation $(12)\sigma$ runs through $S_n \setminus A_n$. Hence the subtrahend $\sum_{\sigma \in S_n \setminus A_n} \alpha_{1,1\sigma}\alpha_{2,2\sigma}\alpha_{3,3\sigma}\cdots\alpha_{n,n\sigma}$ in (i) is equal to

$$\sum_{\sigma \in A_n} \alpha_{1,1(12)\sigma}\alpha_{2,2(12)\sigma}\alpha_{3,3(12)\sigma}\cdots\alpha_{n,n(12)\sigma}$$

$$= \sum_{\sigma \in A_n} \alpha_{1,2\sigma}\alpha_{2,1\sigma}\alpha_{3,3\sigma}\cdots\alpha_{n,n\sigma}$$

$$= \sum_{\sigma \in A_n} \alpha_{2,1\sigma}\alpha_{1,2\sigma}\alpha_{3,3\sigma}\cdots\alpha_{n,n\sigma} \text{ (commutativity of multiplication)}$$

$$= \sum_{\sigma \in A_n} \alpha_{1,1\sigma}\alpha_{2,2\sigma}\alpha_{3,3\sigma}\cdots\alpha_{n,n\sigma} \quad \text{(first two rows are identical)},$$

which is the minuend in (i). Hence $det\ A = 0$. □

It will be convenient to identify the i-th row

$$\alpha_{i1}\ \alpha_{i2} \cdots \alpha_{in}$$

of a matrix $A = (\alpha_{ij}) \in Mat_n(K)$, where K is a field, with the vector

$$(\alpha_{i1}\ \alpha_{i2} \cdots \alpha_{in})$$

in $K^n = Mat_{1\times n}(K)$. Similarly, the j-th column

$$\begin{matrix}\alpha_{1j}\\ \alpha_{2j}\\ \vdots\\ \alpha_{nj}\end{matrix}$$

of A will be identified with the vector (matrix)

$$\begin{pmatrix}\alpha_{1j}\\ \alpha_{2j}\\ \vdots\\ \alpha_{nj}\end{pmatrix}$$

in $Mat_{n\times 1}(K)$. Thus it is meaningful to speak of K-linear (in)dependence of rows and columns of a matrix. Likewise, we can add two rows (columns) and multiply them by scalars.

44.10 Lemma: *Let K be a field and $A,B \in Mat_n(K)$, $n \geqslant 2$. Suppose that B is obtained from A by multiplying a particular row (column) of A by some $\gamma \in K$ and adding it to a different row (column) of A. Then $det\ B = det\ A$.* (The determinant does not change when we add a multiple of a row (column) to another.)

Proof: We prove the assertion about rows only. Suppose that the k-th row in A is multiplied by $\gamma \in K$ and added to the m-th row. Writing $A = (\alpha_{ij})$, $B = (\beta_{ij})$, we have $\beta_{mj} = \gamma\alpha_{kj} + \alpha_{mj}$ and $\beta_{ij} = \alpha_{ij}$ for $i \neq m$. Lemma 44.5 gives $det\ B = det\ C + det\ A$, where $C \in Mat_n(K)$ is identical with A except for the m-th row, which is γ times the k-th row of A. By Lemma 44.4, $det\ C = \gamma det\ D$, where $D \in Mat_n(K)$ is identical with A except that the m-th row of D = the k-th row of A = the k-th row of D. Then $det\ D = 0$ by Lemma 44.9 and $det\ B = \gamma \cdot det\ D + det\ A = det\ A$. □

44.11 Lemma: *Let K be a field and $A \in Mat_n(K)$. If every entry in a particular row (column) of A is equal to $0 \in K$, then $\det A = 0$.*

Proof: Let $A = (\alpha_{ij})$. Under the hypothesis of the lemma, each summand $\varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma}$ $(\sigma \in S_n)$ of $\det A$ is zero, for one of the factors is zero. Hence $\det A = 0$. □

44.12 Lemma: *Let K be a field and $A \in Mat_n(K)$. If the rows (columns) of A are linearly dependent over K, then $\det A = 0$.*

Proof: When $n = 1$, A must be the matrix (0)(see Example 42.2(a)), and $\det A = 0$. Assume now $n \geqslant 2$. We prove the assertion about rows only. If the rows of A are K-linearly dependent, then there are $\beta_1, \beta_2, \ldots, \beta_n$ in K such that

$$\beta_1(\text{1st row}) + \beta_2(\text{2nd row}) + \ldots + \beta_n(n\text{-th row}) = (0,0,\ldots,0)$$

and not all of $\beta_1, \beta_2, \ldots, \beta_n$ are equal to $0 \in K$. Suppose $\beta_k \neq 0$. Then β_k has an inverse β_k^{-1} in K. We multiply the i-th row by $\beta_i\beta_k^{-1}$ and add the product to the kth row; we do this for each $i \neq k$. Then we obtain a matrix B whose determinant is equal to $\det A$ by Lemma 44.10. On the other hand, the k-th row of B consists entirely of zeroes and $\det B = 0$ by Lemma 44.11. Hence $\det A = 0$. □

Now we want to discuss the calculation of determinants. In practice, determinants are almost never computed from the definition

$$\sum_{\sigma \in S_n} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma}$$

Rather, a determinant of an $n \times n$ matrix is expressed in terms of the determinants of certain $(n-1) \times (n-1)$ matrices, these in turn in terms of the determinants of certain $(n-2) \times (n-2)$ matrices and so on, until we come to 2×2 matrices, whose determinants are evaluated readily. This reduction process is known as the expansion of a determinant along (or by) a row (column). To describe this process, we introduce a definition.

44.13 Definition: Let K be a field and $A = (\alpha_{ij}) \in Mat_n(K)$, with $n \geqslant 2$. Let M_{ij} be the $(n-1) \times (n-1)$ matrix obtained from A by deleting the i-th row and the j-th column of A, which intersect at the entry α_{ij} of A. Then $(-1)^{i+j} det\ M_{ij}$ is called the *cofactor of* α_{ij} *in* A. We write A_{ij} for the cofactor of α_{ij} in A.

The following lemma justifies the terminology.

44.14 Lemma: *Let K be a field and $A = (\alpha_{ij}) \in Mat_n(K)$, where $n \geqslant 2$. Let k,m be fixed elements of $\{1,2,\ldots,n\}$. Collecting together all terms containing α_{km} in*

$$det\ A = \sum_{\sigma \in S_n} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma}$$

we write

$$det\ A = \alpha_{km}c_{km} + \text{terms not containing } \alpha_{km}.$$

The c_{km} having been defined uniquely in this way, we claim:

(1) $c_{nn} = $ *cofactor of* $\alpha_{nn} = A_{nn}$,
(2) $c_{nm} = A_{nm}$ *for any* $m = 1,2,\ldots,n$,
(3) $c_{km} = A_{km}$ *for any* $k,m = 1,2,\ldots,n$.

Proof: (1) We have

$$\begin{aligned} det\ A &= \sum_{\sigma \in S_n} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma} \\ &= \sum_{\substack{\sigma \in S_n \\ n\sigma = n}} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n-1,(n-1)\sigma}\alpha_{n,n\sigma} + \sum_{\substack{\sigma \in S_n \\ n\sigma \neq n}} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma} \\ &= \alpha_{nn} \sum_{\substack{\sigma \in S_n \\ n\sigma = n}} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n-1,(n-1)\sigma} + \text{terms not involving } \alpha_{nn}. \end{aligned}$$

Any $\sigma \in S_n$ with $n\sigma = n$ can be regarded as a permutation in S_{n-1}, and any permutation in S_{n-1} can be regarded as a permutation in S_n with $n\sigma = n$. Here $\varepsilon(\sigma)$ is independent of whether we regard σ as an element of S_n or of S_{n-1}. Hence

$$c_{nn} = \sum_{\substack{\sigma \in S_n \\ n\sigma = n}} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n-1,(n-1)\sigma} = \sum_{\sigma \in S_{n-1}} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n-1,(n-1)\sigma}$$

$$= \begin{vmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1,n-1} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2,n-1} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n-1,1} & \alpha_{n-1,2} & \cdots & \alpha_{n-1,n-1} \end{vmatrix} = A_{nn}.$$

(2) We prove $c_{nm} = A_{nm}$ for all $m = 1,2, \ldots, n$. The case $m = n$ having been settled in part (1) above, we assume $m < n$. Consider the matrix $B =$

$$\begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1,m-1} & \alpha_{1n} & \alpha_{1,m+1} & \cdots & \alpha_{1,n-1} & \alpha_{1m} \\ \alpha_{21} & \cdots & \alpha_{2,m-1} & \alpha_{2n} & \alpha_{2,m+1} & \cdots & \alpha_{2,n-1} & \alpha_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{n-1,1} & \cdots & \alpha_{n-1,m-1} & \alpha_{n-1,n} & \alpha_{n-1,m+1} & \cdots & \alpha_{n-1,n-1} & \alpha_{n-1,m} \\ \alpha_{n1} & \cdots & \alpha_{n,m-1} & \alpha_{nn} & \alpha_{n,m+1} & \cdots & \alpha_{n,n-1} & \alpha_{nm} \end{pmatrix}$$

obtained from A by interchanging the m-th and n-th columns. Then we have $det\ A = -\ det\ B$ by Lemma 44.7 and, by part (1),

$$det\ B = \alpha_{nm}\ det\ M\ +\ \text{terms not involving } \alpha_{nm}. \qquad \text{(ii)}$$

where M is the $(n-1) \times (n-1)$ matrix we obtain from B by deleting its n-th row and n-th column. A glance at B reveals that M is obtained from

$$M_{nm} = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1,m-1} & \alpha_{1,m+1} & \cdots & \alpha_{1,n-1} & \alpha_{1n} \\ \alpha_{21} & \cdots & \alpha_{2,m-1} & \alpha_{2,m+1} & \cdots & \alpha_{2,n-1} & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{n-1,1} & \cdots & \alpha_{n-1,m-1} & \alpha_{n-1,m+1} & \cdots & \alpha_{n-1,n-1} & \alpha_{n-1,n} \end{pmatrix}$$

by $n - 1 - m$ interchanges of columns. Hence $det\ M = (-1)^{n-1-m} det\ M_{nm} = -(-1)^{n+m} det\ M_{nm} = -A_{nm}$. Substituting this in (ii), we get

$$\begin{aligned} det\ A = -\ det\ B &= \alpha_{nm}(-\ det\ M)\ -\ \text{terms not involving } \alpha_{nm} \\ &= \alpha_{nm}A_{nm}\ +\ \text{terms not involving } \alpha_{nm}, \end{aligned}$$

as was to be shown.

(3) We now prove $c_{km} = A_{km}$ for all k,m. The case $k = n$ having been settled in part (2), we assume $k < n$. We consider the matrix C obtained

from B by interchanging the k-th and n-th rows. Then $det\ C = -\ det\ B = det\ A$ by Lemma 44.7 and, by part (1),

$$det\ C = \alpha_{km}\ det\ N\ +\ \text{terms not involving } \alpha_{km}$$

where N is the $(n-1) \times (n-1)$ matrix we obtain from C by deleting its n-th row and n-th column. The matrix N is obtained from M_{km} by $n - m - 1$ interchanges of columns and $n - k - 1$ interchanges of rows. Hence

$$det\ N = (-1)^{(n-m-1)+(n-k-1)} det\ M_{km} = (-1)^{k+m} det\ M_{km} = A_{km}$$

and

$$det\ A = det\ C = \alpha_{km}A_{km}\ +\ \text{terms not involving } \alpha_{km}.$$

This completes the proof. □

44.15 Theorem: *Let K be a field, $A = (\alpha_{ij}) \in Mat_n(K)$, where $n \geqslant 2$. Let A_{ij} be the cofactor of α_{ij} in A. Then*

$$det\ A = \alpha_{i1}A_{i1} + \alpha_{i2}A_{i2} + \cdots + \alpha_{in}A_{in}$$
$$det\ A = \alpha_{1j}A_{1j} + \alpha_{2j}A_{2j} + \cdots + \alpha_{nj}A_{nj}$$

for all i,j.

Proof: We have $det\ A = \sum_{\sigma \in S_n} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma}$

$$= \sum_{\substack{\sigma \in S_n \\ i\sigma = 1}} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma} + \sum_{\substack{\sigma \in S_n \\ i\sigma = 2}} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma} + \cdots$$

$$+ \sum_{\substack{\sigma \in S_n \\ i\sigma = n}} \varepsilon(\sigma)\alpha_{1,1\sigma}\alpha_{2,2\sigma}\cdots\alpha_{n,n\sigma}$$

$$= \alpha_{i1}c_{i1} + \alpha_{i2}c_{i2} + \cdots + \alpha_{in}c_{in}$$
$$= \alpha_{i1}A_{i1} + \alpha_{i2}A_{i2} + \cdots + \alpha_{in}A_{in}$$

for any i. This proves the first formula. Applying it with A^t, j in place of A, i, we obtain the second formula. □

The first formula in Theorem 44.15 is known as the *expansion of det A along the i-th row*, the second as the *expansion of det A along the j-th column*. Each element in the i-th row (j-th column) contributes a term,

more specifically α_{ij} contributes $\alpha_{ij}A_{ij}$, where A_{ij} is the determinant of the $(n-1) \times (n-1)$ matrix obtained from A by deleting the row and column of α_{ij}, times 1 or -1, determined by the chessboard pattern

$$\begin{pmatrix} + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

The expansion along a row or column is sometimes given as a recursive definition of determinants in terms of determinants of smaller size.

A specific determinant is computed as follows. If a row or column consists of zeroes, the determinant is 0. Otherwise, we choose and fix a row or column. It will be convenient to choose the row (or column) which has the largest number of zeroes. At least one of the entries on the fixed row (column), say β, is distinct from 0. If a column (row) intersects our fixed row (column) at the entry γ, we add $-\gamma\beta^{-1}$ times the column (row) of β to that column (row). We do this for each column (row). This does not change the determinant, but our fixed row (column) will consist entirely of zeroes, except for the entry β. Expanding the determinant along the fixed row (column), we see that the determinant is equal to βD, where D is the new cofactor of β. We repeat the same procedure with the determinant D, and obtain $D = \beta' D'$, say. Then we repeat the same process with D', etc., until we come to a 2×2 or 3×3 determinant which can be computed easily.

44.16 Examples: (a) Let K be a field and $x_1, x_2, \ldots, x_n$ elements in K. We evaluate $det\ (x_i^{j-1}) =$

$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{vmatrix}.$$

This is known as the Vandermonde determinant. Let us denote it by D_n. We add $-x_n$ times the i-th row to the $(i+1)$-st row $(i = 1,2, \ldots, n-1)$. The only nonzero entry in the new last column will be the entry 1 in the 1st row, n-th column. Expanding D_n along the last column, and taking the

factor $x_i - x_n$ in the i-th column of the cofactor outside the determinant sign by Lemma 44.4 ($i = 1,2,\ldots,n-1$), we obtain

$$D_n = (-1)^{n-1}(x_1 - x_n)(x_2 - x_n)\ldots(x_{n-1} - x_n)D_{n-1}.$$

This holds for any n. Thus

$$\begin{aligned} D_n &= (-1)^{n-1}(x_1 - x_n)(x_2 - x_n)\ldots(x_{n-1} - x_n) \times \\ &\quad (-1)^{n-2}(x_1 - x_{n-1})(x_2 - x_{n-1})\ldots(x_{n-2} - x_{n-1})D_{n-2} \\ &= \cdots \\ &= (-1)^{(n-1)+(n-2)+\cdots+1}(x_1 - x_n)(x_2 - x_n)\ldots(x_{n-3} - x_n)(x_{n-2} - x_n)(x_{n-1} - x_n) \\ &\qquad (x_1 - x_{n-1})(x_2 - x_{n-1})\ldots(x_{n-3} - x_{n-1})(x_{n-2} - x_{n-1}) \\ &\qquad (x_1 - x_{n-2})(x_2 - x_{n-2})\ldots(x_{n-3} - x_{n-2}) \\ &\qquad\qquad \ldots\ldots \\ &\qquad (x_1 - x_2). \end{aligned}$$

Changing the sign of the $\binom{n}{2}$ factors on the right hand side and noting that $(n-1) + (n-2) + \cdots + 1 = \binom{n}{2}$, we finally get

$$D_n = \prod_{i>j} (x_i - x_j),$$

the product being over all $\binom{n}{2}$ pairs (i,j), where $i,j = 1,2,\ldots,n$ and $i > j$.

(b) Let K be a field. The determinant of a matrix $(\alpha_{ij}) \in Mat_n(K)$, where $\alpha_{ij} = 0$ whenever $i > j$, which may be written symbolically

$$\begin{vmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1n} \\ & \alpha_{22} & \alpha_{23} & \cdots & \alpha_{2n} \\ & & \alpha_{33} & \cdots & \alpha_{3n} \\ 0 & & & \ddots & \\ & & & & \alpha_{nn} \end{vmatrix}$$

can be evaluated by expanding successively along the first columns. One finds immediately that $|\alpha_{ij}| = \alpha_{11}\alpha_{22}\alpha_{33}\ldots\alpha_{nn}$. Likewise, the determinant

$$\begin{vmatrix} \alpha_{11} & & & & \\ \alpha_{21} & \alpha_{22} & & 0 & \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & & \\ \cdots & \cdots & \cdots & \ddots & \\ \alpha_{n1} & \alpha_{n2} & \alpha_{n3} & \cdots & \alpha_{nn} \end{vmatrix}$$

is evaluated to be $\alpha_{11}\alpha_{22}\alpha_{33}\ldots\alpha_{nn}$. In particular, the determinant of a *diagonal* matrix

$$\begin{pmatrix} \alpha_{11} & & & & \\ & \alpha_{22} & & 0 & \\ & & \alpha_{33} & & \\ 0 & & & \ddots & \\ & & & & \alpha_{nn} \end{pmatrix}$$

is $\alpha_{11}\alpha_{22}\alpha_{33}\ldots\alpha_{nn}$.

What happens if we use the cofactors of the elements in a different row (column) in the expansion along a particular row (column)? We get zero.

44.17 Theorem: *Let K be a field., $A = (\alpha_{ij}) \in Mat_n(K)$, $n \geqslant 2$. Then*

$$\alpha_{i1}A_{k1} + \alpha_{i2}A_{k2} + \cdots + \alpha_{in}A_{kn} = 0$$
$$\alpha_{1j}A_{1m} + \alpha_{2j}A_{2m} + \cdots + \alpha_{nj}A_{nm} = 0$$

whenever $i \neq k$ and $j \neq m$.

Proof: The first (second) sum is the expansion, along the i-th row (j-th column), of *det* B, where B is the matrix obtained from A by replacing the k-th row (m-th column) of A by its i-th row (j-th column). Since two rows (columns) of B are identical, *det* $A = 0$ by Lemma 44.9. The result follows. □

Using *Kronecker's delta*, which is defined by

$$\delta_{rs} = \begin{cases} 1 \text{ if } r = s \\ 0 \text{ if } r \neq s, \end{cases}$$

so that (δ_{ij}) is the identity matrix I in $Mat_n(K)$, Theorem 44.15 and Theorem 44.17 can be written

$$\alpha_{i1}A_{k1} + \alpha_{i2}A_{k2} + \ldots + \alpha_{in}A_{kn} = \delta_{ik} det\ A$$
$$\alpha_{1j}A_{1m} + \alpha_{2j}A_{2m} + \ldots + \alpha_{nj}A_{nm} = \delta_{jm} det\ A.$$

To express these equations more succintly, we introduce a definition.

44.18 Definition: Let K be a field and $A = (\alpha_{ij}) \in Mat_n(K)$, where $n \geqslant 2$. The $n \times n$ matrix obtained from A by replacing the entry α_{ij} by the cofactor A_{ij} of α_{ij} in A is called the *adjoint of* A. Hence

$$\text{the adjoint of } (\alpha_{ij}) = (A_{ij}).$$

Using this terminology, the equations above can be written as matrix equations,

$$A \cdot (\text{adjoint of } A)^t = (det\ A)I$$
$$A^t \cdot (\text{adjoint of } A) = (det\ A)I.$$

Taking the transposes of both sides in the second equation, we obtain

44.19 Theorem: *Let K be a field and $A \in Mat_n(K)$, where $n \geqslant 2$. Then*

$$A \cdot (adjoint\ of\ A)^t = (det\ A)I = (adjoint\ of\ A)^t \cdot A. \qquad \square$$

44.20 Theorem: *Let K be a field and $A \in Mat_n(K)$. Then A is invertible if and only if $det\ A \in K^\times$. If this is the case, the inverse A^{-1} of A is given by the formula*

$$A^{-1} = \frac{1}{det\ A}(adjoint\ of\ A)^t,$$

where $\frac{1}{det\ A}$ *denotes the inverse of* $det\ A$ *in* K.

Proof: If $det\ A = 0$, then $(det\ A)I = 0 \in Mat_n(K)$, hence, by Theorem 44.19, A is a left zero divisor and a right zero divisor in the ring $Mat_n(K)$. From Lemma 29.10, we deduce that A cannot have a left or right inverse.

Otherwise, $det\ A \neq 0$ and $det\ A$ has an inverse $\frac{1}{det\ A}$ in K. If $n = 1$, then $A = (det\ A)$ and $(\frac{1}{det\ A})$ is the inverse of A. If $n \geqslant 2$, we multiply the members of the equations in Theorem 44.19 by $\frac{1}{det\ A}$ and obtain

$$A \cdot \frac{1}{det\ A}(\text{adjoint of } A)^t = I = \frac{1}{det\ A}(\text{adjoint of } A)^t \cdot A.$$

This shows that $\frac{1}{det\ A}(\text{adjoint of } A)^t$ is an inverse of A. So $A \in GL(n,K)$ and, since $GL(n,K)$ is a group, A has a unique inverse. Hence $\frac{1}{det\ A}(\text{adjoint of } A)^t$ is *the* inverse A^{-1} of A. □

The next theorem is another testimony for the use of determinants.

44.21 Theorem: *Let K be a field and $A \in Mat_n(K)$. Then $det\ A = 0$ if and only if the rows (columns) of A are linearly dependent over K.*

Proof: If the rows (columns) of A are linearly dependent over K, then $det\ A = 0$ by Lemma 44.12.

Assume conversely that $det\ A = 0$. Let $A = (\alpha_{ij})$. Let V be an n-dimensional K-vector space and let $\{v_1, v_2, \ldots, v_n\}$ be a K-basis of V. Then the K-linear transformation $T \in L_K(V,V)$, given by

$$v_i T = \sum_{j=1}^{n} \alpha_{ij} v_j$$

has the associated matrix $(\alpha_{ij}) = A$, which is *not* invertible since $det\ A = 0$. So A is not a unit in $Mat_n(K)$ and T is not a unit in $L_K(V,V)$. Thus T is not an isomorphism. From Theorem 42.22, we conclude that T is not one-to-one. Thus $Ker\ T \neq \{0\}$. Let $v \in Ker\ T$, $v \neq 0$. We have

$$v = \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_n v_n$$

for some suitable scalars $\beta_j \in K$. Here not all of β_j are equal to 0, because $v \neq 0$ and $\{v_1, v_2, \ldots, v_n\}$ is a K-basis of V. Then

$$0 = vT = \Big(\sum_{i=1}^{n} \beta_i v_i\Big)T = \sum_{i=1}^{n} \beta_i (v_i T) = \sum_{i=1}^{n} \beta_i \sum_{j=1}^{n} \alpha_{ij} v_j = \sum_{j=1}^{n} \Big(\sum_{i=1}^{n} \beta_i \alpha_{ij}\Big) v_i,$$

so $$\sum_{i=1}^{n} \beta_i \alpha_{ij} = 0 \quad \text{for } j = 1,2, \ldots,n$$

since $\{v_1, v_2, \ldots, v_n\}$ is a K-basis of V. Thus

$$\beta_1(\text{1st row}) + \beta_2(\text{2nd row}) + \cdots + \beta_n(n\text{-th row}) = (0,0, \ldots,0)$$

with scalars $\beta_1, \beta_2, \ldots, \beta_n \in K$ which are not all equal to 0. So the rows of A are K-linearly dependent. Repeating the same argument with A^t, we see that the columns of A, too, are K-linearly dependent. □

We now establish the multiplication rule for determinants:

44.21 Theorem: *Let K be a field, $n \in \mathbb{N}$.*
(1) *$\det(AB) = (\det A)(\det B)$ for all $A,B \in Mat_n(K)$.*
(2) *$\det I = 1$.*
(3) *$\det A^{-1} = (\det A)^{-1}$ for all $A \in GL(n,K)$.*

Proof: (2) That $\det I = 1$ is a special case of the formula for the determinant of a diagonal matrix discused in Example 44.16(b). And (3) follows from (1) and (2): $(\det A^{-1})(\det A) = \det(A^{-1}A) = \det I = 1$.

We prove (1). Let $A = (\alpha_{ij})$, $B = (\beta_{ij})$, $AB = (\gamma_{ij})$, so that $\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ik}\beta_{kj}$ for all i,j. Then $\det(AB) =$

$$\begin{vmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} \end{vmatrix}$$

$$= \begin{vmatrix} \sum_{k_1=1}^{n} \alpha_{1k_1}\beta_{k_1 1} & \sum_{k_2=1}^{n} \alpha_{1k_2}\beta_{k_2 2} & \cdots & \sum_{k_n=1}^{n} \alpha_{1k_n}\beta_{k_n n} \\ \sum_{k_1=1}^{n} \alpha_{2k_1}\beta_{k_1 1} & \sum_{k_2=1}^{n} \alpha_{2k_2}\beta_{k_2 2} & \cdots & \sum_{k_n=1}^{n} \alpha_{2k_n}\beta_{k_n n} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{k_1=1}^{n} \alpha_{nk_1}\beta_{k_1 1} & \sum_{k_2=1}^{n} \alpha_{nk_2}\beta_{k_2 2} & \cdots & \sum_{k_n=1}^{n} \alpha_{nk_n}\beta_{k_n n} \end{vmatrix}$$

$$= \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \cdots \sum_{k_n=1}^{n} \begin{vmatrix} \alpha_{1k_1}\beta_{k_1 1} & \alpha_{1k_2}\beta_{k_2 2} & \cdots & \alpha_{1k_n}\beta_{k_n n} \\ \alpha_{2k_1}\beta_{k_1 1} & \alpha_{2k_2}\beta_{k_2 2} & \cdots & \alpha_{2k_n}\beta_{k_n n} \\ & & \cdots\cdots & \\ \alpha_{nk_1}\beta_{k_1 1} & \alpha_{nk_2}\beta_{k_2 2} & \cdots & \alpha_{nk_n}\beta_{k_n n} \end{vmatrix} \qquad \text{(Lemma 44.5)}$$

$$= \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \cdots \sum_{k_n=1}^{n} \beta_{k_1 1}\beta_{k_2 2}\cdots\beta_{k_n n} \begin{vmatrix} \alpha_{1k_1} & \alpha_{1k_2} & \cdots & \alpha_{1k_n} \\ \alpha_{2k_1} & \alpha_{2k_2} & \cdots & \alpha_{2k_n} \\ & & \cdots\cdots & \\ \alpha_{nk_1} & \alpha_{nk_2} & \cdots & \alpha_{nk_n} \end{vmatrix} \qquad \text{(Lemma 44.4).}$$

In this n-fold sum, $k_1, k_2, \ldots, k_n$ run independently over $1,2, \ldots, n$. If, however, any two of $k_1, k_2, \ldots, k_n$ are equal, then the determinant $|\alpha_{ik_j}|$ in the n-fold sum has two identical columns and therefore vanishes (Lemma 44.9). So we may disregard those combinations of the indices $k_1, k_2, \ldots, k_n$ which contain two equal values, and restrict the n-fold summation to those combinations of $k_1, k_2, \ldots, k_n$ such that $k_1, k_2, \ldots, k_n$ are all distinct. Then the n-fold sum becomes

$$\sum_{\begin{pmatrix} 1 & 2 & \cdots & n \\ k_1 & k_2 & \cdots & k_n \end{pmatrix} \in S_n} \beta_{k_1 1}\beta_{k_2 2}\cdots\beta_{k_n n} \begin{vmatrix} \alpha_{1k_1} & \alpha_{1k_2} & \cdots & \alpha_{1k_n} \\ \alpha_{2k_1} & \alpha_{2k_2} & \cdots & \alpha_{2k_n} \\ & & \cdots\cdots & \\ \alpha_{nk_1} & \alpha_{nk_2} & \cdots & \alpha_{nk_n} \end{vmatrix}$$

$$= \sum_{\sigma \in S_n} \varepsilon(\sigma)\beta_{1\sigma,1}\beta_{2\sigma,2}\cdots\beta_{n\sigma,n} \cdot \varepsilon(\sigma) \begin{vmatrix} \alpha_{1,1\sigma} & \alpha_{1,2\sigma} & \cdots & \alpha_{1,n\sigma} \\ \alpha_{2,1\sigma} & \alpha_{2,2\sigma} & \cdots & \alpha_{2,n\sigma} \\ & & \cdots\cdots & \\ \alpha_{n,1\sigma} & \alpha_{n,2\sigma} & \cdots & \alpha_{n,n\sigma} \end{vmatrix}$$

$$= \sum_{\sigma \in S_n} \varepsilon(\sigma)\beta_{1\sigma,1}\beta_{2\sigma,2}\cdots\beta_{n\sigma,n} \cdot \begin{vmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ & & \cdots\cdots & \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} \end{vmatrix} \qquad \text{(Lemma 44.8)}$$

$$= \sum_{\sigma \in S_n} \varepsilon(\sigma)\beta_{1\sigma,1}\beta_{2\sigma,2}\cdots\beta_{n\sigma,n} \quad (det\ A) = (det\ A)(det\ B) \qquad \text{(Lemma 44.3).}$$

Hence $det\ AB = (det\ A)(det\ B)$. □

The equation $det\ AB = (det\ A)(det\ B)$ may also be written in the forms

$$det\ AB = (det\ A)(det\ B^t),$$
$$det\ AB = (det\ A^t)(det\ B),$$
$$det\ AB = (det\ A^t)(det\ B^t).$$

So there are four versions of the multiplication rule for determinants, known as the rows by columns multiplication, rows by rows multiplication, columns by columns multiplication, column by rows multiplication, which are respectively described below:

If K is a field, $(\alpha_{ij}), (\beta_{ij}), (\gamma_{ij}) \in Mat_{n\times m}(K)$, and if

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ik}\beta_{kj} \qquad \text{for all } i,j, \qquad \text{or}$$

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ik}\beta_{jk} \qquad \text{for all } i,j, \qquad \text{or}$$

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ki}\beta_{kj} \qquad \text{for all } i,j, \qquad \text{or}$$

$$\gamma_{ij} = \sum_{k=1}^{n} \alpha_{ki}\beta_{jk} \qquad \text{for all } i,j,$$

then $|\gamma_{ij}| = |\alpha_{ij}|\,|\beta_{ij}|$.

Restricting the mapping $det: Mat_n(K) \to K$ to

$$GL(n,K) = \{A \in Mat_n(K): det\ A \in K^\times\}$$

(Theorem 44.20), we obtain a group homomorphism

$$det: GL(n,K) \to K^\times.$$

The kernel

$$\{A \in Mat_n(K): det\ A = 1\}$$

of this determinant homomorphism is a normal subgroup of $GL(n,K)$, known as the *special linear group of degree n over K*, and denoted as $SL(n,K)$.

Exercises

1. Verify that the determinant of a 3×3 matrix (α_{ij}) can be computed as follows. We write the first column of the matrix to the right of the

matrix and the second column to the right of the last written copy of the first column:

$$\begin{matrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{21} & \alpha_{22} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{31} & \alpha_{32} \end{matrix}$$

We take the products of the upper-left, lower-right diagonals (full lines) unchanged, the products of the lower-right, upper-left diagonals (broken lines) with a minus sign. The sum of these six products is the determinant of (α_{ij}). (This rule cannot be extented to $n \times n$ matrices if n is greater than 3).

2. Compute the determinants of the following matrices over $\mathbb{Q}$:

(a) $\begin{pmatrix} 1 & 3 & 5 \\ 0 & 1 & 6 \\ 0 & 0 & 2 \end{pmatrix}$; (b) $\begin{pmatrix} 1 & 4 & 5 \\ -1 & 1 & 0 \\ 1 & 2 & 2 \end{pmatrix}$; (c) $\begin{pmatrix} 1 & 2 & 5 \\ -1 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}$; (d) $\begin{pmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 3 & -4 & 2 \end{pmatrix}$;

(e) $\begin{pmatrix} 2 & 1 & 1 \\ 3 & 1 & 0 \\ 0 & 3 & 2 \end{pmatrix}$; (f) $\begin{pmatrix} 1 & 1 & -1 & -2 \\ 0 & 2 & -3 & 0 \\ 1 & -1 & 0 & 1 \\ 0 & 1 & 0 & 4 \end{pmatrix}$; (g) $\begin{pmatrix} 1 & 0 & 0 & 7 & -3 \\ 4 & 4 & 0 & 1 & 0 \\ 2 & -8 & 2 & -1 & 4 \\ 2 & 1 & 5 & -2 & -1 \\ -1 & 0 & 1 & 0 & 2 \end{pmatrix}$.

3. Find *det A* if A is the matrix $\begin{pmatrix} \bar{2} & \bar{1} & \bar{1} \\ \bar{3} & \bar{1} & \bar{0} \\ \bar{0} & \bar{3} & \bar{2} \end{pmatrix}$ in $Mat_3(\mathbb{Z}_7)$.

4. Expand along the third column:

$$\begin{vmatrix} 1 & 0 & 5 & 4 & 0 \\ 3 & -2 & -1 & 3 & 1 \\ 0 & -3 & 1 & 2 & 0 \\ 1 & 2 & 0 & 2 & 4 \\ 6 & 1 & 2 & 1 & -1 \end{vmatrix}.$$

5. Find the adjoints of $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ -2 & 0 & 4 \end{pmatrix}$ and $\begin{pmatrix} -1 & 2 & 1 & 0 \\ 0 & 1 & 4 & 3 \\ 2 & -2 & 0 & 1 \\ 5 & 1 & 6 & 2 \end{pmatrix}$.

6. Find the inverses of the following matrices:

(a) $\begin{pmatrix} 1 & 0 & 2 \\ -2 & 3 & 1 \\ -3 & 4 & -1 \end{pmatrix}$ over $\mathbb{Q}$; (b) $\begin{pmatrix} 0 & 1 & 3 \\ 5 & 4 & -2 \\ 2 & -1 & 1 \end{pmatrix}$ over $\mathbb{Q}$; (c) $\begin{pmatrix} \bar{1} & \bar{2} & \bar{0} \\ \bar{2} & \bar{1} & \bar{0} \\ \bar{0} & \bar{2} & \bar{2} \end{pmatrix}$ over $\mathbb{Z}_3$;

(d) $\begin{pmatrix} 1 & 0 & 7 & 2 & 3 \\ 6 & 9 & 8 & 1 & 6 \\ 8 & 5 & 9 & 4 & 10 \\ 1 & 0 & 3 & 2 & 7 \\ 2 & 7 & 1 & 5 & 4 \end{pmatrix}$ over $\mathbb{Z}_{11}$.

7. Let K be a field and $n \geqslant 2$. Prove that

$$\text{adjoint of (adjoint of } A) = (\det A)^{n-2}A$$

for any $A \in Mat_n(K)$.

8. Let K be a field and $n \geqslant 2$. Let $x,y \in K$ and put

$$d_n = \begin{vmatrix} x+y & xy & 0 & 0 & \dots \\ 0 & x+y & xy & 0 & \dots \\ 0 & 0 & x+y & xy & \dots \\ & & \dots\dots\dots & & \end{vmatrix}$$

Express d_n in terms of d_{n-1} and d_{n-2}, and evaluate it in closed form.

9. Evaluate the determinant

$$\begin{vmatrix} \binom{c+m-1}{0} & \binom{c+m}{1} & \binom{c+m+1}{2} & \dots & \binom{c+m+m-1}{m} \\ \binom{c+m}{0} & \binom{c+m+1}{1} & \binom{c+m+2}{2} & \dots & \binom{c+m+m}{m} \\ & & \dots\dots\dots & & \\ \binom{c+m+m-1}{0} & \binom{cm+m}{1} & \binom{c+m+m+1}{2} & \dots & \binom{c+m+m+m-1}{m} \end{vmatrix}.$$

10. Let $q \in \mathbb{N}$ and assume that K is a field of q elements. Using Theorem 44.21, find the orders of the groups $GL(n,K)$ and $SL(n,K)$ (cf. §17, Ex.17).

§45
Linear Equations

Let K be a field and $\alpha_{ij}, \beta_i \in K$ $(i = 1,2,\ldots,m; j = 1,2,\ldots,n)$. We ask if there are elements $x_1, x_2, \ldots, x_n$ in K such that

$$\begin{aligned} \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_n &= \beta_1 \\ \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{2n}x_n &= \beta_2 \\ &\cdots\cdots\cdots\cdots \\ \alpha_{m1}x_1 + \alpha_{m2}x_2 + \cdots + \alpha_{mn}x_n &= \beta_m. \end{aligned} \tag{1}$$

(1) is said to be a *system of linear equations*. We will not treat the general problem here. Our objective is this paragraph is to derive necessary and sufficient conditions for the solvability of (1) in the special case $m = n$. Concerning the case $m \neq n$, we will prove only the following consequence of Theorem 42.21.

45.1 Theorem: *Let K be a field and $\alpha_{ij} \in K$ $(i = 1,2,\ldots,m; j = 1,2,\ldots,n)$. If $n > m$, that is to say, if there are more unknowns than equations in the system*

$$\begin{aligned} \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_n &= 0 \\ \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{2n}x_n &= 0 \\ &\cdots\cdots\cdots\cdots \\ \alpha_{m1}x_1 + \alpha_{m2}x_2 + \cdots + \alpha_{mn}x_n &= 0, \end{aligned} \tag{2}$$

then there are elements $x_1, x_2, \ldots, x_n$ in K, not all of them being zero, which satisfy the system (2).

Proof: Of course $x_1 = x_2 = \cdots = x_n = 0$ is a solution of (2), called the trivial solution. We ask whether nontrivial solutions of (2) exit. The claim is that there does exist nontrivial solutions of (2) when $n > m$.

Let $A = (\alpha_{ij}) \in Mat_n(K)$. Setting $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in Mat_{n\times 1}(K)$ and

$0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in Mat_{m\times 1}(K)$, we may write (2) as a matrix equation:

$$AX = 0.$$

The problem is thus: given $A \in Mat_{m\times n}(K)$, is there a nonzero X in $Mat_{n\times 1}(K)$ such that $AX = 0 \in Mat_{m\times 1}(K)$?

Since $A(N + M) = AN + AM$ and $A(\alpha N) = \alpha(AN)$ for any $N, M \in Mat_{n\times 1}(K)$ and $\alpha \in K$, the mapping

$$\begin{aligned} \varphi: Mat_{n\times 1}(K) &\to Mat_{m\times 1}(K) \\ N &\to AN \end{aligned}$$

is a vector space homomorphism. From Theorem 42.21, we obtain

$$\begin{aligned} n = dim_K Mat_{n\times 1}(K) &= dim_K Ker\,\varphi + dim_K Im\,\varphi \\ &\leqslant dim_K Ker\,\varphi + dim_K Mat_{m\times 1}(K) \\ &= dim_K Ker\,\varphi + m, \end{aligned}$$

so

$$\begin{aligned} dim_K Ker\,\varphi &\geqslant n - m > 0, \\ Ker\,\varphi &\neq \{0\} \subseteq Mat_{n\times 1}(K), \end{aligned}$$

and there does exist an $X \neq 0$ in $Ker\,\varphi$. So there is a nonzero $X \in Mat_{n\times 1}(K)$ with $AX = 0$, as was to be shown. □

45.2 Theorem: *Let K be a field, $(\alpha_{ij}) \in Mat_n(K)$ and let $\beta_1, \beta_2, \ldots, \beta_n$ be elements of K. If $det\,(\alpha_{ij}) \neq 0$, then the system*

$$\begin{aligned} \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_n &= \beta_1 \\ \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{2n}x_n &= \beta_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ \alpha_{n1}x_1 + \alpha_{n2}x_2 + \cdots + \alpha_{nn}x_n &= \beta_n \end{aligned} \tag{3}$$

has a unique solution in K, given by

$$x_j = \frac{det\,B_j}{det\,(\alpha_{ij})}, \qquad (j = 1,2,\ldots,n),$$

where B_j is the matrix obtained from (α_{ij}) by replacing its j-th column by

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}.$$

Proof: Let $A = (\alpha_{ij})$, $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in Mat_{n\times 1}(K)$ and $B = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \in Mat_{n\times 1}(K)$. Then (3) can be written as a matrix equation:

$$AX = B. \tag{4}$$

Multiplying both sides of (4) on the left by $A^{-1} = \frac{1}{det\ A}$ (adjoint of $A)^t$, we obtain

$$X = \frac{1}{det\ A} \text{ (adjoint of } A)^t B. \tag{5}$$

Also, multiplying both sides of (5) on the left by A, and using Theorem 44.12, we obtain (4). Thus (4) and (5) are equivalent. So the system (3) or (4) has a unique solution given by (5). In more detail, when we write A_{ij} for the cofactor of α_{ij} in A, so that (adjoint of A) = (A_{ij}), the solution is given by

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \frac{1}{det\ A} \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

$$= \frac{1}{det\ A} \begin{pmatrix} A_{11}\beta_1 + A_{21}\beta_2 + \cdots + A_{n1}\beta_n \\ A_{12}\beta_1 + A_{22}\beta_2 + \cdots + A_{n2}\beta_n \\ \cdots\cdots\cdots\cdots \\ A_{1n}\beta_1 + A_{2n}\beta_2 + \cdots + A_{nn}\beta_n \end{pmatrix}$$

So $x_j = \frac{1}{det\ A}(\beta_1 A_{1j} + \beta_2 A_{2j} + \cdots + \beta_n A_{nj})$ for $j = 1,2, \ldots, n$. Comparing the expression in parentheses with the expansion

$$\alpha_{1j}A_{1j} + \alpha_{2j}A_{2j} + \cdots + \alpha_{nj}A_{nj}$$

(Theorem 44.15) of $det\ (\alpha_{ij})$ along the j-th column, we see that the parenthetical expression is the expansion, along the j-th column, of the de-

terminant of the matrix B_j that is obtained from (α_{ij}) by replacing its j-th column by B. Thus

$$x_j = \frac{det\ B_j}{det\ (\alpha_{ij})} \qquad (j = 1,2,\ldots,n),$$

as claimed. □

The formula $x_j = \dfrac{det\ B_j}{det\ (\alpha_{ij})}$ is known as Cramer's rule after G. Cramer (1704-1752).

45.3 Theorem: *Let K be a field, $(\alpha_{ij}) \in Mat_n(K)$. The system*

$$\begin{aligned} \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_n &= 0 \\ \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{2n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ \alpha_{n1}x_1 + \alpha_{n2}x_2 + \cdots + \alpha_{nn}x_n &= 0 \end{aligned} \tag{6}$$

has a nontrivial solution in K (i.e., a solution distinct from the obvious one $x_1 = x_2 = \ldots = x_n = 0$), if and only if $det\ (\alpha_{ij}) = 0$.

Proof: If $det\ (\alpha_{ij}) \neq 0$, then the system has a unique solution by Theorem 45.2, which must be $x_1 = x_2 = \cdots = x_n = 0$, as follows also from Cramer's rule, for the numerator determinants, having a column consisting of zeroes only, are all equal to 0. Thus, if the system has a nontrivial solution in K, then $det\ (\alpha_{ij})$ must be zero.

Suppose conversely that $det\ (\alpha_{ij}) = 0$. Then the columns of (α_{ij}) are linearly dependent over K (Theorem 44.21): There are elements $\beta_1, \beta_2, \ldots, \beta_n$ in K, not all of them being zero, such that

$$\beta_1\begin{pmatrix}\alpha_{11}\\ \alpha_{21}\\ \vdots\\ \alpha_{n1}\end{pmatrix} + \beta_2\begin{pmatrix}\alpha_{12}\\ \alpha_{22}\\ \vdots\\ \alpha_{n2}\end{pmatrix} + \cdots + \beta_n\begin{pmatrix}\alpha_{1n}\\ \alpha_{2n}\\ \vdots\\ \alpha_{nn}\end{pmatrix} = \begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix}.$$

Thus $x_1 = \beta_1, x_2 = \beta_2, \ldots, x_n = \beta_n$ is a nontrivial solution of (6). □

45.4 Remark: The theorems in this paragraph are chiefly of theoretical interest. Finding solutions of specific systems by the methods described in this paragraph would be very tedious.

Exercises

1. Find all solutions of the following systems of linear equations:

(a)
$$\begin{aligned} 3x + 4y - 5z &= -1 \\ 2x - 3y + z &= 3 \\ 2x + y + 6z &= 0; \end{aligned}$$

(b)
$$\begin{aligned} 4x + y - 5z - u &= 1 \\ 6x + 2y - 3z + 3u &= 8 \\ -4x + 5y - 2z + u &= -3 \\ 2x - 7z - 3u &= 0. \end{aligned}$$

2. Using Cramer's rule, find the solutions in $\mathbb{Z}_{13}$ of the following systems of linear equations, where $\bar{\ }$ denotes residue classes modulo 13:

(a)
$$\begin{aligned} \bar{2}x + \bar{11}y + \bar{4}z &= \bar{1} \\ \bar{3}x + \bar{8}y + \bar{5}z &= \bar{6} \\ \bar{9}x + \bar{12}y + \bar{4}z &= \bar{7}; \end{aligned}$$

(b)
$$\begin{aligned} \bar{2}x + \bar{11}y + \bar{4}z + \bar{3}u &= \bar{5} \\ \bar{8}x + \bar{10}y + \bar{6}z + \bar{7}u &= \bar{2} \\ \bar{1}x + \bar{9}y + \bar{2}z + \bar{8}u &= \bar{6} \\ \bar{3}x + \bar{1}y + \bar{0}z + \bar{5}u &= \bar{4}. \end{aligned}$$

§46
Algebras

In this last paragraph of Chapter 4, we consider multiplication of vectors. If, on a vector space, there is an associative multiplication which is distributive over addition and compatible with multiplication by scalars, the vector space is said to be an algebra. The formal is definition is as follows.

46.1 Definition: Let K be a field and $(V,+)$ an abelian group. A quintuple $(V,+,\circ,K,\cdot)$ is called an *algebra over* K, or a *K-algebra* provided

(i) $(V,+,\circ)$ is a ring,

(ii) $(V,+,K,\cdot)$ is a vector space,

(iii) $(\alpha\cdot a)\circ b = \alpha\cdot(a\circ b) = a\circ(\alpha\cdot b)$ for all $\alpha \in K$, $a,b \in V$.

It is implicit in this definition that $\circ$ is a binary operation on V, called multiplication, and $\cdot$ is a mapping from $K \times V$ into V, called multiplication by scalars. As usual, we drop these symbols and write αa for $\alpha\cdot a$, and ab for $a\circ b$. Then (iii) becomes a kind of associativity law: $(\alpha a)b = \alpha(ab) = a(\alpha b)$. As usual, we shall call V, rather than the quintuple $(V,+,\circ,K,\cdot)$, a K-algebra.

Examples: (a) Let K be a field and L a field containing K. Then L is a algebra over K.

(b) Let K be a field. Then $Mat_n(K)$ is a K-vector space (Theorem 43.4) and also a ring (Theorem 43.11). Since $(\alpha A)B = \alpha(AB) = A(\alpha B)$ for all α in K and A,B in $Mat_n(K)$ (see (e) in §43, p. 533), we conclude that $Mat_n(K)$ is a K-algebra.

(c) Let K be a field and V a vector space over K. Then $L_K(V,V)$ is a K-vector space (Theorem 43.1) and also a ring (Theorem 43.12). Moreover, whenever $\alpha \in K$ and $T,S \in L_K(V,V)$, there hold

$$v((\alpha T)S) = (v(\alpha T))S = ((\alpha v)T)S = (\alpha v)(TS) = v(\alpha(TS))$$

and

$$v(T(\alpha S)) = (vT)(\alpha S) = (\alpha(vT))S = \alpha((vT)S) = \alpha(v(TS)) = (\alpha v)(TS) = v(\alpha(TS))$$

for all $v \in V$, thus $(\alpha T)S = \alpha(TS) = T(\alpha S)$. Thus $\boldsymbol{L}_K(V,V)$ is a K-algebra.

(d) Let K be a field and x an indeterminate over K. Then $K[x]$ is a vector space over K (Example 39.2(d)) and also a ring. We have $(af(x))g(x) = a(f(x)g(x)) = f(x)(ag(x))$ for all $a \in K$ and $f(x),g(x) \in K[x]$. Thus $K[x]$ is an algebra over K. Likewise the ring $K[x_1,x_2,\ldots,x_n]$ of polynomials in n indeterminates is an algebra over K.

46.3 Lemma: *Let K be a field and V a finite dimensional vector space over K. Suppose there is a multiplication on V which is distributive over addition, and suppose that*

$$(\alpha a)c = \alpha(ac) = a(\alpha c) \qquad \text{for all } \alpha \in K \text{ and } a,c \in V$$

(thus all conditions for V to be an algebra over K are satisfied except that associativity of multiplication is open).

Let B be a K-basis of V. Then multiplication on V is associative and V is a K-algebra if and only if

$$(bb')b'' = b(b'b'') \qquad \text{for all } b',b',b'' \in B.$$

Proof: If multiplication on V is associative, then $(bb')b'' = b(b'b'')$ holds for all elements b',b',b'' of V, in particular, for all b',b',b'' in B.

Assume conversely that $(bb')b'' = b(b'b'')$ for all b',b',b'' in B. We put $B = \{b_1,b_2,\ldots,b_n\}$. If x,y,z are arbitrary elements of V, we write them as

$$x = \sum_{i=1}^{n} \alpha_i b_i, \qquad y = \sum_{j=1}^{n} \beta_j b_j, \qquad z = \sum_{k=1}^{n} \gamma_k b_k$$

with suitable scalars $\alpha_i, \beta_j, \gamma_k$. Using distributivity and (iii), we find

$$\begin{aligned}(xy)z &= \Big(\sum_{i=1}^{n} \alpha_i b_i \sum_{j=1}^{n} \beta_j b_j\Big) \cdot \sum_{k=1}^{n} \gamma_k b_k = \sum_{i,j=1}^{n} (\alpha_i b_i)(\beta_j b_j) \cdot \sum_{k=1}^{n} \gamma_k b_k \\ &= \sum_{i,j=1}^{n} \alpha_i (b_i(\beta_j b_j)) \cdot \sum_{k=1}^{n} \gamma_k b_k = \sum_{i,j=1}^{n} \alpha_i (\beta_j (b_i b_j)) \cdot \sum_{k=1}^{n} \gamma_k b_k \\ &= \sum_{i,j=1}^{n} (\alpha_i \beta_j)(b_i b_j) \cdot \sum_{k=1}^{n} \gamma_k b_k = \sum_{i,j,k=1}^{n} ((\alpha_i \beta_j)(b_i b_j))(\gamma_k b_k)\end{aligned}$$

$$= \sum_{i,j,k=1}^{n} (\alpha_i\beta_j)((b_ib_j)(\gamma_kb_k)) \qquad = \sum_{i,j,k=1}^{n} (\alpha_i\beta_j)[\gamma_k((b_ib_j)b_k)]$$

$$= \sum_{i,j,k=1}^{n} ((\alpha_i\beta_j)\gamma_k)((b_ib_j)b_k)$$

and likewise $x(yz) = \sum_{i=1}^{n} \alpha_i b_i \cdot \left(\sum_{j=1}^{n} \beta_j b_j \sum_{k=1}^{n} \gamma_k b_k\right) = \sum_{i=1}^{n} \alpha_i b_i \cdot \sum_{j,k=1}^{n} (\beta_j\gamma_k)(b_jb_k)$

$$= \sum_{i,j,k=1}^{n} (\alpha_i(\beta_j\gamma_k))(b_i(b_jb_k)).$$

Now $(\alpha_i\beta_j)\gamma_k = \alpha_i(\beta_j\gamma_k)$ since the multiplication on K is associative and $(b_ib_j)b_k = b_i(b_jb_k)$ by hypothesis, so $(xy)z = x(yz)$. Hence the multiplication on V is also associative. □

46.4 Examples: (a) Let K be a field and G a finite multiplicative group. Let KG denote the K-vector space that has G as a K-basis. Thus the elements of KG are sums $\sum_{i=1}^{|G|} \alpha_i g_i$, where $G = \{g_1, g_2, \ldots, g_{|G|}\}$. It will be convenient to modify this notation as $\sum_{g\in G} \alpha_g g$. Two elements $\sum_{g\in G} \alpha_g g$ and $\sum_{g\in G} \beta_g g$ of KG are equal if and only if $\alpha_g = \beta_g$ for each $g \in G$. The sum of $\sum_{g\in G} \alpha_g g$ and $\sum_{g\in G} \beta_g g$ is $\sum_{g\in G} (\alpha_g + \beta_g)g$, and the product of $\gamma \in K$ by $\sum_{g\in G} \alpha_g g$ is $\sum_{g\in G} \gamma\alpha_g g$. We now define a multiplication on KG by extending the multiplication on G using distributivity. More precisely, we define the product of $\sum_{g\in G} \alpha_g g$ by $\sum_{g\in G} \beta_g g = \sum_{h\in G} \beta_h h$ to be $\sum_{g\in G} \alpha_g g \sum_{h\in G} \beta_h h =$

$$\sum_{g,h\in G} \alpha_g\beta_h gh = \sum_{k\in G} \Big(\sum_{\substack{g,h\in G \\ gh=k}} \alpha_g\beta_h \Big) k.$$

For any $a = \sum_{g\in G} \alpha_g g$, $b = \sum_{g\in G} \beta_g g$ in KG and $\gamma \in K$, we have $(\gamma a)b =$

$$\Big(\gamma \sum_{g\in G} \alpha_g g\Big)\sum_{g\in G} \beta_g g = \sum_{g\in G} \gamma\alpha_g g = \sum_{g\in G} \beta_g g = \sum_{k\in G}\Big(\sum_{\substack{g,h\in G\\ gh=k}} \gamma\alpha_g\beta_h\Big)k =$$

$$\sum_{k\in G} \gamma\Big(\sum_{\substack{g,h\in G\\ gh=k}} \alpha_g\beta_h\Big)k = \gamma\sum_{k\in G}\Big(\sum_{\substack{g,h\in G\\ gh=k}} \alpha_g\beta_h\Big)k = \gamma(ab) \text{ and similarly } a(\gamma b) = \gamma(ab).$$

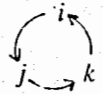
The reader will verify that distributivity laws are valid.

Each element g_0 of G can be regarded as an element $\sum_{g\in G} \alpha_g g$ in KG, where $\alpha_g = 0$ if $g \neq g_0$ and $\alpha_{g_0} = 1$. Thus we regard G as a subset of KG. It is checked easily that, for any $g,h \in G$, the product of gh in KG is the product gh in G. Since multiplication on G is associative, and since G is a K-basis of KG, we learn from Lemma 46.3 that KG is an algebra over K. It is called the *group algebra of G over K*.

(b) Let $\mathbb{H} = \mathbb{R}^4$ be the four-dimensional $\mathbb{R}$-vector space of ordered quadruples, and let $e = (1,0,0,0)$, $i = (0,1,0,0)$, $j = (0,0,1,0)$, $k = (0,0,0,1)$. Thus $\{e,i,j,k\}$ is a basis of $\mathbb{H}$ over $\mathbb{R}$. We give a multiplication table for these basis elements:

	e	i	j	k
e	e	i	j	k
i	i	$-e$	k	$-j$
j	j	$-k$	$-e$	i
k	k	j	$-i$	$-e$

Thus $ea = ae = a$ for any $a \in \{i,j,k\}$ and the products of i,j,k are like the cross product of the vectors i,j,k in $\mathbb{R}^3$. The product of two distinct elements from $\{i,j,k\}$ is equal to $\mp$ the third, the sign being "+" for products taken in the order indicated in the accompanying diagram, and "-" in the reverse order.



By distributivity, we have the product formula:

$$(\alpha e + \beta i + \gamma j + \delta k)(\alpha' e + \beta' i + \gamma' j + \delta' k) = (\alpha\alpha' - \beta\beta' - \gamma\gamma' - \delta\delta')e + (\alpha\beta' + \beta\alpha' + \gamma\delta' - \delta\gamma')i + (\alpha\gamma' - \beta\delta' + \gamma\alpha' + \delta\beta')j + (\alpha\delta' + \beta\gamma' - \gamma\beta' + \delta\alpha')k$$

which may be taken as the definition of multiplication on $\mathbb{H}$. One checks that this multiplication is distributivite over addition, and that e is an identity element. To prove the associativity of multiplication, we must only verify the $4^3 = 64$ equations $(ab)c = a(bc)$, where $a,b,c \in \{e,i,j,k\}$ (Lemma 46.3). This verification is left to the reader. The multiplication is thus seen to be associative. One also finds immediately $(\alpha a)b = \alpha(ab) = a(\alpha b)$ for any $\alpha \in \mathbb{R}$ and $a,b \in \mathbb{H}$. Thus $\mathbb{H}$ is an algebra over $\mathbb{R}$. This algebra was discovered by the Irish mathematician W. R. Hamilton (1805-1865). The elements of $\mathbb{H}$ are called *quaternions*, and $\mathbb{H}$ is known as the *Hamilonian algebra of quaternions*. It is not commutative, since $ij = e \neq -e = ji$, for example.

Since e is the identity of $\mathbb{H}$, we will write 1 instead of e and α instead of αe (here $\alpha \in \mathbb{R}$). Then any real number α can be thought of as a quaternion $\alpha 1 = \alpha + 0i + 0j + 0k$. In like manner, any complex number $\alpha + \beta i$ (where $\alpha,\beta \in \mathbb{R}$) can be considered as a quaternion $\alpha + \beta i + 0j + 0k$. In this way, we may suppose that $\mathbb{R}$ and $\mathbb{C}$ are subrings of $\mathbb{H}$.

For any $a \in \mathbb{H}$, say $\alpha + \beta i + \gamma j + \delta k$ with $\alpha,\beta,\gamma,\delta \in \mathbb{R}$, we say α is the *real part of* a and $\beta i + \gamma j + \delta k$ is the *imaginary part of* a. We also put $\bar{a} = \alpha - \beta i - \gamma j - \delta k$ and call $\bar{a}$ the *conjugate of* a. It is easily seen that $\overline{ab} = \bar{b}\,\bar{a}$ for any $a,b \in \mathbb{H}$ (note the reversal of the conjugates). We define the *norm of* a, denoted as $N(a)$, to be $a\bar{a}$. Thus $N(\alpha + \beta i + \gamma j + \delta k)$ is equal to $\alpha^2 + \beta^2 + \gamma^2 + \delta^2$. Note that $N(a) \in \mathbb{R}$. Clearly $N(a) = 0$ if and only if $a = 0$.

There holds $N(ab) = ab\,\overline{ab} = ab\,\bar{b}\,\bar{a} = aN(b)\bar{a} = N(b)a\,\bar{a} = N(b)N(a) = N(ab)$ for any quaternions $a,b \in \mathbb{H}$. This is equivalent to the identity

$$(\alpha^2 + \beta^2 + \gamma^2 + \delta^2)(\alpha'^2 + \beta'^2 + \gamma'^2 + \delta'^2) = (\alpha\alpha' - \beta\beta' - \gamma\gamma' - \delta\delta')^2 + (\alpha\beta' + \beta\alpha' + \gamma\delta' - \delta\gamma')^2 + (\alpha\gamma' - \beta\delta' + \gamma\alpha' + \delta\beta')^2 + (\alpha\delta' + \beta\gamma' - \gamma\beta' + \delta\alpha')^2$$

which holds in fact in any commutative ring. Thus the product of two numbers, each of which is a sum of four squares, is also a sum of four squares.

Just like we divide a complex number $a = \alpha + \beta i$ by a nonzero complex number $b = \gamma + \delta i$ by multiplying the numerator and denominator of a/b by the conjugate $\bar{b} = \gamma - \delta i$ of b:

$$\frac{a}{b} = \frac{\alpha + \beta i}{\gamma + \delta i} = \frac{\alpha + \beta i}{\gamma + \delta i}\frac{\gamma - \delta i}{\gamma - \delta i} = \frac{\alpha\gamma + \beta\delta}{\gamma^2 + \delta^2} + \frac{-\alpha\delta + \beta\gamma}{\gamma^2 + \delta^2} i,$$

we can divide any quaternion a by any nonzero quaternion b by multiplying the "numerator" and "denominator" of a/b by the conjugate $\bar{b}$:

$$\frac{a}{b} = \frac{a\bar{b}}{b\bar{b}} = \frac{a\bar{b}}{N(b)}.$$

More exactly, any nonzero quaternion b has a multiplicative inverse $(1/N(b))\bar{b}$. Thus $\mathbb{H}$ is a division ring. An algebra which is a division ring is called a division algebra. So $\mathbb{H}$ is a division algebra.

An interesting theorem of F. G. Frobenius (1849-1917) states that $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{H}$ are the only finite dimensional division algebras over $\mathbb{R}$.

(c) The last example can be generalized. Let K be a field in which $1 + 1$ is distinct from 0. Let $Q = K^4$ be the four-dimensional K-vector space of ordered quadruples, and let $e = (1,0,0,0)$, $i = (0,1,0,0)$, $j = (0,0,1,0)$, $k = (0,0,0,1)$. Thus $\{e,i,j,k\}$ is a basis of Q over K. We define a multiplication on Q by

$$\begin{aligned}(\alpha e + \beta i + \gamma j + \delta k)(\alpha' e + \beta' i + \gamma' j + \delta' k) &= (\alpha\alpha' - \beta\beta' - \gamma\gamma' - \delta\delta')e \\ &+ (\alpha\beta' + \beta\alpha' + \gamma\delta' - \delta\gamma')i \\ &+ (\alpha\gamma' - \beta\delta' + \gamma\alpha' + \delta\beta')j \\ &+ (\alpha\delta' + \beta\gamma' - \gamma\beta' + \delta\alpha')k\end{aligned}$$

This multiplication is associative, distributivite over addition and e is an identity element. One checks easily $(\alpha a)b = \alpha(ab) = a(\alpha b)$ for any $\alpha \in K$ and $a,b \in Q$. Thus Q is a K-algebra. Q is called the *algebra of quaternions over* K. This time it will be convenient *not* to identify $\alpha \in K$ with $\alpha e \in Q$.

The *conjugate* $\bar{a}$ *of* $a = \alpha e + \beta i + \gamma j + \delta k \in Q$ is defined to be $\alpha e - \beta i - \gamma j - \delta k$ and the *norm* $N(a)$ *of* a *to be* $a\bar{a}$. Thus $N(\alpha e + \beta i + \gamma j + \delta k) = \alpha^2 + \beta^2 + \gamma^2 + \delta^2$. If K is a field such that $\alpha^2 + \beta^2 + \gamma^2 + \delta^2 = 0$ implies $\alpha = \beta = \gamma = \delta = 0$, then any nonzero $a \in Q$ has a multiplicative inverse $(1/N(a))\bar{a}$ and Q is a division algebra. Otherwise, Q has zero divisors: there is a nonzero $a \in Q$ such that $a\bar{a} = 0$.

Exercises

1. Multiply $2\iota + 3(12) + (13) - 2(23) + (123) - 3(132)$ by $\iota + 2(12) + 4(13) - 3(23) + 2(123) + (132)$ in $\mathbb{Q}S_3$.

2. Let G be a finite group, K a field. Put $e = \sum_{g \in G} g \in KG$. Show that $e^2 = |G|e$.

3. Let K be a field and A an algebra over K. Prove that the center $Z(A)$ of A (see §32, Ex. 1) is a subspace of A.

4. Let G be a finite group. Show that $dim_{\mathbb{Q}} Z(\mathbb{Q}G)$ is equal to the number of conjugacy classes in G.

5. For any $a \in \mathbb{H}$, show that there are real numbers t,n such that
$$a^2 - ta + n = 0.$$

6. Prove that $a^2ia^2 + ia^2ia - iaia^2 - aia^2i = 0$ for any $a \in \mathbb{H}$.

7. Let $a,b \in \mathbb{H}$. Show that $ab = ba$ if and only if $1,a,b$ are linearly dependent over $\mathbb{R}$.

8. Prove that $\{\mp 1, \mp i, \mp j, \mp k\} \subseteq \mathbb{H}$ is a group isomorphic to Q_8 (see §17, Ex.15) and that $S = \{\mp 1, \mp i, \mp j, \mp k, \frac{\mp 1 \mp i \mp j \mp k}{2}\} \subseteq \mathbb{H}$ is a group isomorphic to $SL(2,\mathbb{Z}_3)$. Show that $\{\mp 1\} \trianglelefteq S$ and $S/\{\mp 1\} \cong A_4$.

9. Prove that the quaternion algebra over $\mathbb{C}$ is isomorphic (as ring and $\mathbb{C}$-vector space) to the $\mathbb{C}$-algebra $Mat_2(\mathbb{C})$.

10. Let K be a field in which $1 + 1 \neq 0$ and α,β nonzero elements in K. Let A be the four dimensional K-vector space with K-basis e,i,j,k. On A, we define a multiplication by the multiplication table on the basis elements:

	e	i	j	k
e	e	i	j	k
i	i	αe	k	j
j	j	$-k$	βe	$-\beta i$
k	k	$-\alpha j$	βi	$-\alpha\beta e$

(a) Prove that this multiplication makes A into a K-algebra ($\mathbb{H}$ is a special case $K = \mathbb{R}$, $\alpha = \beta = -1$).

(b) Show that the center of A is $\{ke \in A : k \in K\}$ and that A has no ideals aside from 0 and A.

(c) Define the *conjugate* $\bar{a}$ *of* $a = \alpha e + \beta i + \gamma j + \delta k \in A$ to be $\alpha e - \beta i - \gamma j - \delta k$ and the *norm* $N(a)$ *of* a *to be* $a\bar{a}$. Verify $\overline{ab} = \bar{b}\,\bar{a}$ and $N(ab) = N(a)N(b)$ for any $a,b \in A$.

(d) Prove that A is a division algebra if and only if $N(a) \neq 0$ for any nonzero $a \in A$ and this holds if and only if $\gamma_0^2 = \alpha\gamma_1^2 + \beta\gamma_2^2$ implies $\gamma_0 = \gamma_1 = \gamma_2 = 0$ for any $\gamma_0, \gamma_1, \gamma_2 \in K$.

(e) If K is finite, say $|K| = q$, show that there are $q + 1$ elements in $\{\alpha\gamma_1^2 \in K : \gamma_1 \in K\}$ and $\{1 - \beta\gamma_2^2 \in K : \gamma_2 \in K\}$ and conclude that A is not a division algebra (This is a special case of an important theorem due to H. J. M. Wedderburn (1882-1948) which states that any finite division algebra is a field).

11. If $1 + 1 = 0$ in a field K and A is as in Ex.10, show that the mapping $x \to x^2$ is a ring homomorphism from A into A.

CHAPTER 5

Fields

§47
Historical Introduction

For a long time in the history of mathematics, algebra was understood to be the study of roots of polynomials.

This must be clearly distinguished from numerical computation of the roots of a given specific polynomial. The Newton-Hörner method is the best known procedure to evaluate roots of polynomials. The actual calculation of roots was (and is) a minor point. The principal object of algebra was understanding the structure of the roots: how they depend on the coefficients, whether they can be given in a formula, etc.

There is, of course, the related question concerning the existence of roots of polynomials. Does every polynomial have a root? Here the coefficient of polynomials were implicitly understood to be real numbers. A. Girard (1595-1632) expressed that any polynomial has a root in some realm of numbers (not neccessarily in the realm of complex numbers), without indicating any method of proof. R. Descartes (1596-1650) noted that $x - c$ is a divisor of a polynomial if c is a root of that polynomial and gave a rule for determining the number of real roots in a specified interval. He makes an obscure remark about the existence of roots. Euler stated that any polynomial has a root in complex numbers. This result came to be called the fundamental theorem of algebra, a very inappropriate name.

Euler proved it rigorously for polynomials of degree $\leqslant 6$. J. R. D'Alembert (1717-1783), Lagrange, P. S. Laplace (1749-1827) made attempts to prove this statement. As Gauss criticized, their proof actually assumes the existence of a root in some realm of numbers, and shows that the root is in fact in $\mathbb{C}$. Gauss himself gave several proofs, some of which cannot be accepted as rigorous by modern standards. Nevertheless, Gauss has the credit for having given the first valid demonstration of the so-called fundamental theorem of algebra. After Kronecker established in 1882 that any polynomial has a root is some realm of "numbers" (see §51), the earlier attempts became rigorous proofs. The really fundamental theorem is Kronecker's theorem.

This assures the existence of roots, but does not bring insight to the problem of understanding the nature of roots any more than existence theorems about differential equations give solutions of differential equations or information about their analytic bahavior, singularities, asymptotic expansions, etc.

The solution of quadratic equations were known to many ancient civilizations. The cubic and biquadratic polynomials (that is, polynomials of degree four) were treated by Italian mathematicians. Scipione del Ferro (1465-1526) succeeded in solving the cubic equation $x^3 + ax = b$ (1515) in terms of radical expresions. In 1535, Tartaglia (1499/1500-1557) solved the cubic polynomial of the form $x^3 + ax^2 = b$. G. Cardan (1501-1576), substituted $x - (b/3)$ for x and transformed the general cubic $x^3 + bx^2 + cx + d$ to a form in which the x^2 term is absent. Thus assuming, with no loss of generality, the equation to be $x^3 + px + q = 0$, a formula for the roots is found to be

$$\sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}$$

This is known as Cardan's formula, but it is actually Tartaglia who found it and divulged it to Cardan under pledge of secrecy, who later broke his promise and published it in his book *Ars Magna* (1545). Cardan's originality lay in reducing the general cubic to one of the form $x^3 + px + q$, discussing the so called irreducible case, noting that a cubic can have at most three roots and making an introduction to the theory of symmetric polynomials.

Cardan's book contains a method for finding roots of biquadratic polynomials (that is, polynomials of degree four) discovered by his pupil L. Ferrari (1522-1565) round 1540. This book made a great impact on the developement of algebra. Cardan even calculated with complex numbers, which manifested themselves to be indispensable. Contrary to what one may be at first inclined to believe, there was no need for complex numbers as far as quadratic equations are concerned: mathematicians had declared such equations as $x^2 = -1$ simply unsolvable. However, in Cardan's formula, one has to take square roots of negative numbers even if all the roots are real (the irreducible case). In fact, the roots of a cubic polynomial whose three roots are real cannot be expressed by a formula involving real radicals only (Lemma 59.30).

Thus the first half of 16th century witnessed remarkable achievements in algebra. As late as 1494, Fra Luca Pacioli had expressed that a cubic equation cannot be solved by radicals, and by 1540 both the cubic and the biquadratic equation was solved by radicals. The next step would be to find a formula for the roots of a quintic polynomial (that is, polynomials of degree five) and, better still, of a polynomial of n-th degree in general.

Other solutions of polynomial equations of the degree $\leqslant 4$ are later given by Descartes, Walter von Tschirnhausen (ca. 1690) and Euler. Noted mathematicians tried in vein to find a formula for the roots of a quintic polynomial. Mathematicians began to suspect that a quintic polynomial equation cannot be solved by radicals.

Lagrange published in 1770-1771 a long paper "Réflexions sur la résolution algébrique des équations" in which he studied extensively all known methods of solutions of polynomial equations. His aim was to derive a general procedure from the known methods for finding roots of polynomials. He treated quadratic, cubic and biquadratic polynomials in detail, and succeeded in subsuming the various methods under one general principle. The roots of a polynomial are expressed in terms of a quantity t, called the *resolvent*, and the resolvent t itself is the root of an auxiliary polynomial, called the *resolvent polynomial*. When the degree of the given polynomial is n, the resolvent polynomial is of degree $(n-1)!$ in x^n. For $n \leqslant 4$, the auxiliary equation has therefore a smaller degree than the polynomial given, and can be solved algebraically (by

induction), but for $n \geqslant 5$, solving the auxiliary equation is not easier than to solving the original equation.

The resolvent is a function of the roots which is invariant under some but not all of the permutation of the roots. For example, when the degree is four, $r_1r_2 + r_3r_4$ does not change if the roots r_1,r_2 and r_3,r_4 are interchanged. Lagrange is thus led to the permutation of the roots, i.e., he investigated, without appropriate terminology and notation, the symmetric group on n letters. (Incidentally, the degree of the resolvent polynomial is a divisor of $n!$, the order of the symmetric group. This is how Lagrange came to Theorem 10.9.)

Lagrange noted that, in the successful cases $n \leqslant 4$, the resolvent has the form $r_1 + \alpha r_2 + \cdots + \alpha^{n-1}r_n$, where r_i are the roots of the polynomial and α is a root of $x^n - 1$. This type of a resolvent does not work in case $n = 5$, but it is concievable that expressions of some other kind could work as resolvents. Lagrange studied which type of expressions could be resolvents.

In 1799, P. Ruffini (1765-1822) claimed a proof of the impossibility of solving the general quintic equation algebraically, but whether his proof was rigorous remained controversial. In 1826, Abel gave the first complete proof of this impossibility theorem. His proof consists of two parts. In the first part, he found the general form of resolvents must be as in Lagrange's description for the cubic and biquadratic cases; in the second part, he demonstrated that it can never be a root of a polynomial of fifth degree. He added, without proof, that the general equation cannot be solved algebraically if the degree is greater than 5. In addition to the general equation, Abel also investigated which special equations can be solved by radicals. He proved a theorem which reads, in modern terminology, that an equation is solvable by radicals if the associated Galois group is commutative. It is in this connection that commutative groups are called abelian.

Abel thus finally demonstrated that the *general* equation cannot be solved by radicals. "General polynomial" means that the coefficients are independent variables or, more in the spirit of algebra, indeterminates. Abel's theorem does not say anything about polynomials whose coefficients are fixed complex numbers. But some polynomial equations with constant coefficients of degree five or greater *are* solvable by

radicals. What is the criterion for a polynomial equation to be solvable by radicals? This question is resolved by the French mathematician Évariste Galois (1811-1832). With Galois, the principle subject matter of algebra definitely ceased to be polynomial equations. Galois marks the beginning of modern algebra, which means the study of algebraic structures (groups, rings, vector spaces, fields, and many others).

*

* *

Galois had a short and dramatic life. He began publishing articles when he was a pupil in Lycée (1828). He was a remarkable talent and a difficult student. He wanted to enter the École Polytechnique, but failed twice in the entrance examinations. The reason, he says later, was that the questions were so simple that he refused to answer them. He later entered the École Normale (1829), but expelled from it due to a letter in the student newspaper. His unbearable pride was notorious. He became politicized, was sent to jail for some months, then began a liason with "une coquette de bas étage" and died in an obscure duel (1830).

Galois' achievements have not been appreciated by his contemporaries. He submitted several papers to the French Academy, but these were rejected as unintelligible. It was not until J. Liouville (1809-1882) published his memoirs in 1846 that the world came to know Galois and realize him to be the one of the greatest mathematicians of all time.

Galois associated, with each resolvent equation, a field intermediate between the field of the coefficients of the polynomial and the field of the roots. His ingenious idea is to associate, with the given polynomial and intermediate fields, a series of groups and to translate assertions about fields into group-theoretical statements. This involved the clarification of the field and group concepts. The theory of groups is founded by Galois. He proved that a polynomial equation is solvable by radicals if and only if, in the series of groups, each group is normal and of prime index in the next one, i.e., if and only if the group of the polynomial is solvable in the sense of Definition 27.19 (Theorem 27.25).

It should be noted that this criterion is not an effective procedure to determine actually whether a polynomial equation is solvable by

radicals. His contemporaries expected that "the condition of solvability, if it exists, ought to have an external character which can be verified by inspecting the coefficients of a given equation or, all the better, by solving other equations of degrees lower than that of the equation to be solved."[1] His is not a workable test that effectively decides if an equation is solvable by radicals. Galois himself writes: "If now you give me an equation that you have chosen at pleasure, and if you want to know if it is or if it is not solvable by radicals, I need do nothing more than indicate to you the means of answering your question, without wanting to give myself or anyone else the task of doing it. In a word, the calculations are impractical."[2] But this is the whole point. Who cares about solvability of polynomial equations. What Galois achieved, and what his contemporaries failed to appreciate, is a fascinating parallel between the group and field structures. The group-theoretical solvability condition is at best a trivial application of the theory.

This was too big a change in algebra and in mathematics and heralded the end of an era when mathematics was the science of numbers and figures. Ever since the time of Gauss and Galois, mathematics is the science of structures. Galois theory is the first mathematical theory that compares two different structures: fields and groups. It was not easy to follow this developement. Even mathematicians of later generations concieved Galois theory as a tool for answering certain questions in the theory of equations. The first writer on Galois theory who clearly differentiated between the theory and its applications is Heinrich Weber (1842-1913). In his famous text-book on algebra (1894), the exposition of the theory occupies one chapter, its applications another.

The first writer on Galois theory is E. Betti (1823-1892). He published a paper "Sulla risoluzione delle equazioni algebriche" in 1852, in which he closely follows Galois' line. This is more of a commentary than an original exposition. Here the concept of conjugacy and of factor groups made a appeared dimly. Another commentator on Galois theory is J. A. Serret (1819-1885).

Camile Jordan (1838-1922) gave the first exposition of Galois theory that does not follow Galois' own line. With Jordan, emphasis shifted from polynomials to groups. He made many important original contributions. Among other things, he clarified the relationship between irreducible

polynomials and transitive groups, developed the theory of transitive groups, defined factor groups as the group of the auxiliary equation, introduced composition series, proved that the composition factors in any two composition series of a solvable group are isomorphic. The group concept became central, but solving polynomial equations still remained as the major concern.

At the same time, Two German mathematicians, L. Kronecker and R. Dedekind (1831-1916), were making very significant contributions to field theory.

Dedekind lectured on Galois theory as early as 1856. He seems to be the first mathematician who realized that the Galois group should be regarded as an automorphism group of a field rather than a group of permutations. In fact, he uses the term "permutation" for what we now call a field automorphism. This means, of course, he very rightly recognized the theory as a theory on fields, not as a theory on polynomials. He introduced the notion of dependence/independence of elements in an extension field over the base field.

Kronecker discussed adjunction in detail. He noted that it is possible to adjoin transcendental elements as well as algebraic ones to a field and proved the important theorem that any polynomial splits into linear factors in some extension field.

Weber carried Kronecker's and Dedekind's ideas further. His exposition, the first modern treatment of the subject, is not restricted to $\mathbb{Q}$, but rather deals with an arbitrary field. He clearly states that the theory is about field extensions and automorphism groups of these extensions. Weber was far ahead of his time. Many mathematicians of his time found his treatment abstract and difficult.

Then came Emil Artin (1898-1962). He combined techniques of linear algebra and field theory. Extensions are sometimes regarded as fields, sometimes as vector spaces, whichever may be convenient. He studied automorphisms of fields, proved that the degree of an extension is equal to the order of the automorphism group, introduced the notion of a Galois extension, and abolished the role of the resolvent (primitive element). This latter was an ugly aspect of the theory about field extensions, remnant of earlier times when the theory has been regarded

as one about polynomials. Artin then set up the correspondence between intermediate fields and subgroups of the automorphism group. All computations are eliminated from the theory. Where an earlier writer would spend many pages for the step-by-step adjunction of resolvents to construct a splitting field, we see Artin merely write: "Let E be a splitting field of $f(x)$." With Artin, Galois theory lost all its connections with its past. It is interesting to note that Artin does to applications of the theory to polynomial equations. In his book *Galois Theory*, applications are harshly separated from the main text: they can be no more than an appendix; but Artin does not even condescend to write the appendix himself: this task is relegated to one of his students.

[1] Poisson, quoted from Kiernan's article (see References), page 76.
[2] Galois, quoted from Edwards' book *Galois Theory*, page 81.

§48
Field Extensions

We recall a technical term from Example 39.2(f).

48.1 Definition: Let E be a field and let K be a nonempty subset of E. If K itself is a field under the operations defined on E, then K is called a *subfield of* E. In this case, E is called an *extension field of* K, or simply an *extension of* K.

We write E/K to denote that E is an extension of K, and speak of the field extension E/K. Confusion with a factor group or a factor space is not likely. We will frequently employ Hasse diagrams (see §21) for field extensions. For example, the picture

will mean that K is a subfield of E.

As in the case of subgroups, subrings and subspaces, we have a subfield criterion.

48.2 Lemma (Subfield criterion): *Let E be a field and K a nonempty subset of E. Then K is a subfield of E if and only if*

(i) $a + b \in K$,
(ii) $-b \in K$,
(iii) $ab \in K$,
(iv) $b^{-1} \in K$ (*in case* $b \neq 0$)

for all $a, b \in K$.

Proof: A field is a ring in which the nonzero elements form a commutative group under multiplication (see the remarks after Definition 29.13). Thus E is a ring of this type, and K is a subfield of E if and only if K is a

subring of E such that the nonzero elements in K form a commutative group under multiplication. Certainly, every subgroup of $E^\times = E\backslash\{0\}$ is commutative. Thus K is a subfield of E if and only if K is a subring of E and $K\backslash\{0\}$ is a subgroup of $E^\times$. Now K is a subring of E if and only if (i),(ii),(iii) hold and $K\backslash\{0\}$ is a subgroup of $E^\times$ if and only if

(iii)′ $ab \in K\backslash\{0\}$ for all $a,b \in K\backslash\{0\}$

and (iv) hold. Since $K \subseteq E$ and the field E has no zero divisors, (iii)′ is weaker than (iii), and we conclude that K is a subfield of E if and only if (i),(ii),(iii),(iv) hold. □

From now on, we will write $\frac{1}{b}$ (or $1/b$) for the inverse b^{-1} of a nonzero element in a field. Likewise, we will write $\frac{a}{b}$ or (a/b) for the product $ab^{-1} = b^{-1}a$ of two elements a,b^{-1} in a field (assuming $b \neq 0$). It follows from Lemma 48.2 that, whenever K is a subfield of E and $a,b \in K$, then

$$a + b,\ a - b,\ ab,\ \frac{a}{b}$$

belong to K, it being assumed $b \neq 0$ in the last case. A subfield of E is therefore a nonempty subset of E that is closed under addition, subtraction, multiplication and division (by nonzero elements).

48.3 Examples: (a) $\mathbb{R}$ is a extension of $\mathbb{Q}$, and $\mathbb{C}$ is an extension of $\mathbb{Q}$. Also $\mathbb{R}$ is a subfield of $\mathbb{C}$.

(b) If K is any field and x an indeterminate over K, then K is a subfield of $K(x)$ (provided we identify, as usual, an element a of K with the rational function $\frac{a}{1}$, where the numerator and denominator are elements of $K \subseteq K[x]$). Similarly K is a subfield of $K(x,y)$, where y is another indeterminate over K.

(c) Let $\mathbb{Q}(i) := \{x + yi \in \mathbb{C} : x,y \in \mathbb{Q}\} \subseteq \mathbb{C}$. For any a,b in $\mathbb{Q}(i)$, say $a = x + yi$ and $b = z + ui$ with $x,y,z,u \in \mathbb{Q}$, we have

(i) $a + b = (x + z) + (y + u)i \in \mathbb{Q}(i)$,

(ii) $-b = (-z) + (-u)i \in \mathbb{Q}(i)$,

(iii) $ab = (xz - yu) + (xu + yz)i \in \mathbb{Q}(i)$,

(iv) $b^{-1} = \frac{z}{z^2 + u^2} + \frac{-u}{z^2 + u^2} i \in \mathbb{Q}(i)$, provided $b = z + ui \neq 0 + 0i = 0$.

So $\mathbb{Q}(i)$ is a subfield of $\mathbb{C}$. It is in fact the field of fractions of $\mathbb{Z}[i]$, and is called the *gaussian field.*

(d) $\mathbb{Q}(\sqrt{2}) := \{x + y\sqrt{2} \in \mathbb{R} : x,y \in \mathbb{Q}\}$ is a subfield of $\mathbb{R}$. Indeed, for any a,b in $\mathbb{Q}(\sqrt{2})$, say $a = x + y\sqrt{2}$ and $b = z + u\sqrt{2}$ with $x,y,z,u \in \mathbb{Q}$, we have

(i) $a + b = (x + z) + (y + u)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$,

(ii) $-b = (-z) + (-u)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$,

(iii) $ab = (xz + 2yu) + (xu + yz)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$,

(iv) $b^{-1} = \frac{z}{z^2 - 2u^2} + \frac{-u}{z^2 - 2u^2}\sqrt{2} \in \mathbb{Q}(\sqrt{2})$, provided $b = z + u\sqrt{2} \neq 0 + 0\sqrt{2} = 0$. Here we use the fact that $\sqrt{2} \in \mathbb{R}$ is an irrational number (Example 35.11) so that $z^2 - 2u^2 \neq 0$ if z and u are nonzero rational numbers.

(e) Let $L = \{x + y\sqrt[3]{2} \in \mathbb{R} : x,y \in \mathbb{Q}\} \subseteq \mathbb{R}$. Then L is not a subfield of $\mathbb{R}$ since, for example $\sqrt[3]{2} \in L$ but $\sqrt[3]{2}\cdot\sqrt[3]{2} \notin L$ (why?) On the other hand, $\mathbb{Q}(\sqrt[3]{2}) := \{x + y\sqrt[3]{2} + z\sqrt[3]{4} \in \mathbb{R} : x,y,z \in \mathbb{Q}\} = \{x + y\sqrt[3]{2} + z(\sqrt[3]{2})^2 \in \mathbb{R} : x,y,z \in \mathbb{Q}\}$ is a subfield of $\mathbb{R}$. The proof of $b \in \mathbb{Q}(\sqrt[3]{2})\backslash\{0\} \implies 1/b \in \mathbb{Q}(\sqrt[3]{2})$ is left to the reader.

(f) Let K be a field and let K_i $(i \in I)$ be a family of subfields of K. Then $\bigcap_{i \in I} K_i$ is a subfield of K, for the closure properties in Lemma 48.2 hold for $\bigcap_{i \in I} K_i$ if they hold for each of the K_i.

From the last example, we infer that the intersection of *all* subfields of a field K is a subfield of K. Note that the intersection is taken over a nonempty set, since at least K is a subfield of K.

48.4 Definition: Let K be a field. The intersection of all subfields of K is called the *prime subfield of* K.

Thus every subfield of K contains (is an extension of) the prime subfield of K. We want to describe the elements in the prime subfield of K. Let P denote the prime subfield of K. In order to distinguish clearly between the integer $1 \in \mathbb{Z}$ and the identity element of K, we will denote in this discussion the identity element of K as e. We know $0 \in P$, $e \in P$ and $0 \neq e$ because P is a field. Now P is a group under addition, so $e + e = 2e$, $2e + e = 3e$, $3e + e = 4e$, ... are elements of P, and also $-e, -2e, -3e, -4e, \ldots$.

Hence $\ldots, -4e, -3e, -2e, -e, 0, e, 2e, 3e, 4e, \ldots$

all belong to P: we have $\{me \in K : m \in \mathbb{Z}\} \subseteq P$. Moreoever, P is closed under division (by nonzero elements), and so $P_0 := \{me/ne \in K : m,n \in \mathbb{Z}\}$ is a subset of P. It is natural to expect that P_0 is a subfield of K (and thus $P_0 = P$): for any me/ne, $re/se \in P_0$ with $m,n,r,s \in \mathbb{Z}$, we presumably have

(i) $\dfrac{me}{ne} + \dfrac{re}{se} = \dfrac{(ms + rn)e}{(ns)e} \in P_0$,

(ii) $-\dfrac{re}{se} = \dfrac{(-r)e}{se} \in P_0$,

(iii) $\dfrac{me}{ne} \dfrac{re}{se} = \dfrac{(mr)e}{(ns)e} \in P_0$,

(iv) $\dfrac{1}{\frac{re}{se}} = \dfrac{se}{re} \in P_0$, provided $\dfrac{re}{se} \neq 0$, i.e., $re \neq 0$.

These are in fact true, but care must be exercised in justifying (i),(ii),(iii), (iv). This is done in the next theorem which states that P is isomorphic either to $\mathbb{Q}$ or to $\mathbb{Z}_p$ for some prime number p.

48.5 Theorem: *The prime subfield of any field K is isomorphic to $\mathbb{Q}$ or to $\mathbb{Z}_p$ for some prime number p (ring isomorphism).*

Proof: Let e be the identity of K and let P be the prime subfield of K. Then $1e = e \neq 0 \neq -e = -1e$. We distinguish two cases, according as there does or does not exist an integer $n \in \mathbb{Z}\backslash\{0\}$ satisfying $ne = 0$.

Case 1. Assume there is a nonzero integer n such that $ne = 0$. Then there are natural numbers k with $ke = 0$. Let p be the smallest natural number such that $pe = 0$. We claim that the mapping

$$\varphi: \mathbb{Z} \to P$$
$$n \to ne$$

is a ring homomorphism, that p is a prime number and that $P \cong \mathbb{Z}_p$.

For any $m,n \in \mathbb{Z}$, we have $(m+n)\varphi = (m+n)e = me + ne$ (this is not distributivity!) and $(mn)\varphi = (mn)e = (me)(ne) = m\varphi \cdot n\varphi$ (here $(mn)e = (me)(ne)$ is distributivity!), so φ is a ring homomorphism.

If p were composite, say $p = rs$ with $r,s \in \mathbb{N}$, $1 < r < p$, $1 < s < p$, then $0 = pe = (rs)e = (re)(se)$ would yield, since the field K has no zero divisors, that $re = 0$ or $se = 0$, contradicting the definition of p as the *smallest* natural number satisfying $pe = 0$. So p is a prime number.

To prove $P \cong \mathbb{Z}_p$, we will find $Ker\ \varphi$. From $pe = 0$, we have $p \in Ker\ \varphi$, so $pn \in Ker\ \varphi$ for all $n \in \mathbb{Z}$ (because $Ker\ \varphi$ is an ideal of $\mathbb{Z}$) and $p\mathbb{Z} \subseteq Ker\ \varphi$. On the other hand, if $m \in Ker\ \varphi$, we divide m by p to get $m = qp + r$, with $q,r \in \mathbb{Z}$ and $0 \leqslant r < p$. This gives $0 = me = (qp + r)e = (qp)e + re = 0 + re$. As $0 \leqslant r < p$, this forces $r = 0$, which means $m = qp$ and $m \in p\mathbb{Z}$. So we get $Ker\ \varphi \subseteq p\mathbb{Z}$. Therefore $Ker\ \varphi = p\mathbb{Z}$. [A more conceptual argument: $Ker\varphi$ is an ideal of $\mathbb{Z}$ and $\mathbb{Z}$ is a principal ideal domain, so $Ker\ \varphi = d\mathbb{Z}$ for some $d \in \mathbb{Z}$. We have $d \neq 0$ in Case 1. From $pe = 0$ we get $p \in Ker\ \varphi = d\mathbb{Z}$, so $d|p$. But p is a prime number, so $d = \mp 1$ or $d = \mp p$. The possibility $d = \mp 1$ is excluded, because $\mp 1e = \mp e \neq 0$. Hence $d = \mp p$ and $Ker\ \varphi = d\mathbb{Z} = \mp p\mathbb{Z} = p\mathbb{Z}$.]

Thus $\mathbb{Z}_p = \mathbb{Z}/p\mathbb{Z} = \mathbb{Z}/Ker\ \varphi \cong Im\ \varphi \subseteq P$ and $Im\ \varphi$, being a ring isomorphic to $\mathbb{Z}_p$, is a field. So $Im\ \varphi$ is a subfield of K, therefore $P \subseteq Im\ \varphi$. This yields $P = Im\ \varphi$ and $\mathbb{Z}_p \cong P$, as claimed.

Case 2. Assume there is no nonzero integer n such that $ne = 0$. We claim that the mapping

$$\psi: \mathbb{Q} \to P$$
$$m/n \to me/ne$$

is a ring homomorphism and that $P \cong \mathbb{Q}$.

First we show that ψ is well defined. If $\frac{m}{n} = \frac{m'}{n'} \in \mathbb{Q}$ with $m,n,m',n' \in \mathbb{Z}$ $(n \neq 0 \neq n')$, then $mn' = m'n$ in $\mathbb{Z}$, so $(mn')e = (m'n)e$ in P, thus $(me)(n'e) = (m'e)(ne)$ in P. Multiplying both sides of this equation by $\frac{1}{ne}\frac{1}{n'e} \in P$, we obtain $\frac{me}{ne} = \frac{m'e}{n'e}$. So ψ is well defined.

ψ is a ring homomorphism: for all $\frac{m}{n}, \frac{r}{s} \in \mathbb{Q}$ with $m,n,r,s \in \mathbb{Z}$, $n \neq 0 \neq s$, we have $\left(\frac{m}{n} + \frac{r}{s}\right)\psi = \left(\frac{ms + rn}{ns}\right)\psi = \frac{(ms + rn)e}{(ns)e} = \frac{(ms)e + (rn)e}{ne \cdot se}$

$$= \frac{(me)(se) + (re)(ne)}{ne \cdot se} = \frac{me}{ne} + \frac{re}{se} = \frac{m}{n}\varphi + \frac{r}{s}\varphi$$

and $\left(\frac{m}{n} \frac{r}{s}\right)\varphi = \frac{mr}{ns}\varphi = \frac{(mr)e}{(ns)e} = \frac{(me)(re)}{(ne)(se)} = \frac{me}{ne}\frac{re}{se} = \frac{m}{n}\varphi\frac{r}{s}\varphi$

Since we assume that $me \neq 0$ for $m \in \mathbb{Z}\backslash\{0\}$ in Case 2, we obtain $Ker\ \varphi = \{\frac{m}{n} \in \mathbb{Q}: \frac{me}{ne} = 0 \in K\} = \{\frac{m}{n} \in \mathbb{Q}: me = 0 \in K\} = \{\frac{m}{n} \in \mathbb{Q}: m = 0 \in \mathbb{Z}\} = \{0\}$, so $\mathbb{Q} \cong \mathbb{Q}/\{0\} = \mathbb{Q}/Ker\ \varphi \cong Im\ \varphi \subseteq P$ and $Im\ \varphi$, being a ring isomorphic to $\mathbb{Q}$, is a field. So $Im\ \varphi$ is a subfield of K, therefore $P \subseteq Im\ \varphi$. This yields $P = Im\ \varphi$ and $\mathbb{Q} \cong P$, as claimed. □

48.6 Definition: Let K be a field and let e be the identity element of K. If there are nonzero integers n such that $ne = 0$, and if p is the smallest natural number such that $pe = 0$, then K is said to be a *field of characteristic* p and p is called the *characteristic of* K. If there is no nonzero integer n such that $ne = 0$, then K is said to be a *field of characteristic* 0, and 0 is called the *characteristic of* K.

Equivalently, K is of characteristic p or 0 according as its prime subfield is isomorphic to $\mathbb{Z}_p$ or to $\mathbb{Q}$. We write $char\ K = p$ and $char\ K = 0$ in these respective cases. For example, $char\ \mathbb{Z}_p = p$ and $char\ \mathbb{Q}(i) = char\ \mathbb{Q}(\sqrt{2}) = char\ \mathbb{R} = char\ \mathbb{C} = 0$. We will usually identify $\mathbb{Z}_p$ or $\mathbb{Q}$ with the prime subfield of K, as the case may be. In particular, we will write 1 instead of e for the identity element of K. Thus K will be considered to be an extension of $\mathbb{Z}_p$ or $\mathbb{Q}$.

We remark that, if K is a field of characteristic p, then $pa = 0$ for any element a of K. This follows from

$$pa = a + a + \cdots + a = 1a + 1a + \cdots + 1a = (1 + 1 + \cdots + 1)a = (p1)a = 0a = 0,$$

the sums having p terms. This result will be used in the sequel without explicit mention.

We make two conventions. Henceforward, we will write $\mathbb{F}_p$ in place of $\mathbb{Z}_p$. This will always remind us that $\mathbb{F}_p$ is a field (p prime). Secondly, we shall drop the bars in the elements of $\mathbb{F}_p$, as we have already done on several

occasions. For example, we will write 2 instead of $\overline{2} \in \mathbb{F}_5$. A notation such as "2" is therefore ambiguous: it stands for the integer $2 \in \mathbb{Z}$, as well as $\overline{2} \in \mathbb{F}_2$, as well as $\overline{2} \in \mathbb{F}_3$, as well as $\overline{2} \in \mathbb{F}_5$, etc. It will be clear from the context, however, which meaning is accorded to "2". The ambiguity is therefore harmless.

We proceed to discuss field homomorphisms.

48.7 Lemma: *If K is a field, then K and $\{0\}$ are the only ideals of K.*

Proof: If A is an ideal of K and $A \neq \{0\}$, there is an $a \in A$, $a \neq 0$. Then a has an inverse $\frac{1}{a}$ in K and $\frac{1}{a}a = 1 \in A$, because A is an ideal. Then we get $b = b \cdot 1 \in A$ for all $b \in K$, so $K \subseteq A$ and $A = K$. □

48.8 Lemma: *If K_1 and K_2 are fields and $\varphi: K_1 \to K_2$ is a ring homomorphism, then either $a\varphi = 0$ for all $a \in K_1$ or φ is one-to-one.*

Proof: $Ker\ \varphi$ is an ideal of K_1, so either $Ker\ \varphi = K_1$ or $Ker\ \varphi = \{0\}$ by Lemma 48.7. In these respective cases, either $a\varphi = 0$ for all $a \in K_1$ or φ is one-to-one. □

When we deal with fields and ring homomorphisms from a field to another, we naturally want to disregard the uninteresting ring homomorphism that maps every element of its domain to the zero element of the other field. Any other ring homomorphism is one-to-one by Lemma 48.8. This leads us to the following definition.

48.9 Definition: If K_1 and K_2 are fields and $\varphi: K_1 \to K_2$ is a one-to-one ring homomorphism, then φ will be called a *field homomorphism*. If φ is a field homomorphism onto K_2, then φ will be called a *field isomorphism*. A field isomorphism from K onto the same field K will be called a *(field) automorphism of K*.

If $\varphi: K_1 \to K_2$ is a field isomorphism, then φ is a homomorphism of additive groups, so $0_{K_1}\varphi = 0_{K_2}$, and also $Ker\ \varphi = \{0_{K_1}\}$, where 0_{K_1} and 0_{K_2} are the zero elements of the fields K_1, K_2, respectively. Thus $\varphi_{K_1\setminus\{0\}}$ is a one-to-one mapping from $K_1\setminus\{0\}$ onto $K_2\setminus\{0\}$. In addition, $(ab)\varphi = a\varphi \cdot b\varphi$ for all a,b in K_1, so $(ab)\varphi = a\varphi \cdot b\varphi$ for all $a,b \in K_1\setminus\{0\}$ and therefore $\varphi_{K_1^\times}: K_1^\times \to K_2^\times$ is a one-to-one homomorphism of groups onto $K_2^\times$: we have $K_1^\times \cong K_2^\times$ In particular, $(1_{K_1})\varphi = 1_{K_2}$, where 1_{K_1} and 1_{K_2} are the identities of the fields K_1, K_2, respectively.

48.10 Lemma: *Let K_1, K_2, K_3 be fields.*

(1) *If $\varphi: K_1 \to K_2$ and $\psi: K_2 \to K_3$ are field homomorphisms, then $\varphi\psi: K_1 \to K_3$ is a field homomorphism.*

(2) *If $\varphi: K_1 \to K_2$ and $\psi: K_2 \to K_3$ are field isomorphisms, then $\varphi\psi: K_1 \to K_3$ is a field isomorphism.*

(3) *If $\varphi: K_1 \to K_2$ is a field isomorphism, then $\varphi^{-1}: K_2 \to K_1$ is a field isomorphism.*

Proof: (1) $\varphi\psi$ is a ring homomorphism by Lemma 30.16(1) and one-to-one by Theorem 3.11(2).

(2) $\varphi\psi$ is a field homomorphism by part (1) and onto by Theorem 3.11(1).

(3) φ^{-1} is a ring homomorphism by Lemma 30.16(2) and one-to-one by Theorem 3.17(1). □

A field homomorphism $\varphi: K_1 \to K_2$ can be characterized as a one-to-one function such that

$$(a+b)\varphi = a\varphi + b\varphi, \qquad (a-b)\varphi = a\varphi - b\varphi, \qquad (ab)\varphi = a\varphi \cdot b\varphi, \qquad \frac{a}{b}\varphi = \frac{a\varphi}{b\varphi}$$

for all $a,b \in K_1$ ($b \neq 0$ in the division). Let us consider some examples.

48.11 Examples: (a) The conjugation mapping $\varphi: \mathbb{C} \to \mathbb{C}$ is an automorphism of $\mathbb{C}$, because $x \to \bar{x}$

$$\overline{x + y} = \bar{x} + \bar{y}, \quad \overline{x - y} = \bar{x} - \bar{y}, \quad \overline{xy} = \bar{x} \cdot \bar{y}, \quad \overline{x/y} = \bar{x}/\bar{y}$$

for any $x,y \in \mathbb{C}$.

(b) The mapping $\varphi: \mathbb{Q}(\sqrt{2}) \to \mathbb{Q}(\sqrt{2})$ is an automorphism of $\mathbb{Q}(\sqrt{2})$ because

$$a + b\sqrt{2} \to a - b\sqrt{2}$$

$$\begin{aligned}((a + b\sqrt{2}) + (c + d\sqrt{2}))\varphi &= ((a + c) + (b + d)\sqrt{2})\varphi = (a + c) - (b + d)\sqrt{2} \\ &= (a - b\sqrt{2}) + (c - d\sqrt{2}) = (a + b\sqrt{2})\varphi + (c + d\sqrt{2})\varphi,\end{aligned}$$

$$\begin{aligned}((a + b\sqrt{2})(c + d\sqrt{2}))\varphi &= ((ac + 2bd) + (ad + bc)\sqrt{2})\varphi \\ &= (ac + 2bd) - (ad + bc)\sqrt{2} = (ac + 2(-b)(-d)) + (a(-d) + (-b)c)\sqrt{2} \\ &= (a - b\sqrt{2})(c - d\sqrt{2}) = (a + b\sqrt{2})\varphi(c + d\sqrt{2})\varphi\end{aligned}$$

for all $a + b\sqrt{2}, c + d\sqrt{2} \in \mathbb{Q}(\sqrt{2})$, where $a,b,c,d \in \mathbb{Q}$, so that φ is a ring homomorphism and, because of $1\varphi = (1 + 0\sqrt{2})\varphi = 1 - 0\sqrt{2} = 1 \neq 0$, the kernel of φ is not K and φ is therefore one-to-one.

(c) Let K be a field and x an indeterminate over K. Then the mapping

$$\varphi: K(x) \to K(x)$$
$$\frac{p(x)}{q(x)} \to \frac{p(x^2)}{q(x^2)}$$

is a field homomorphism. Note that $Im\ \varphi \subset K(x)$. Thus $K(x)$ is isomorphic to a proper subset of itself (namely to $Im\ \varphi$).

Let E/K be a field extension. Then E is an additive group and

$$\begin{aligned}a(x + y) &= ax + ay \\ (a + b)x &= ax + bx \\ (ab)x &= a(bx) \\ 1x &= x\end{aligned}$$

for all $x,y \in E$ and for all $a,b \in K$ (in fact for all $a,b \in E$, but we do not need this now). Hence E is a vector space over K, as we have already noted in Example 39.2(h). Studying both the field and the vector space structure of E will be very useful. In particular, the dimension of E over K will play an important role.

48.12 Definition: Let E/K be a field extension. The dimension of E over K is called the *degree of E over K*, or the *degree of the extension E/K*.

It will prove convenient to write $|E:K|$ instead of $dim_K E$ for the degree of E over K. The field E is said to be a *finite dimensional extension* or *an infinite dimensional extension of K* according as $|E:K|$ is finite or infinite. Most authors use the term "finite extension" for a finite dimensional extension.

An important fact in the theory of fields is that a finite dimensional extension of a finite dimensional extension is a finite dimensional extension, and that the degrees behave multiplicatively. More exactly, we have the

48.13 Theorem: *Let F/E and E/K be field extensions of finite degrees $|F:E|$ and $|E:K|$. Then F/K is a finite dimensional extension. In fact*

$$|F:K| = |F:E|\,|E:K|$$

and furthermore if $\{f_1, f_2, \ldots f_r\}$ is an E-basis of F and $\{e_1, e_2, \ldots, e_s\}$ a K-basis of E, then $\{f_i e_j : i = 1,2,\ldots,r;\ j = 1,2,\ldots,s\}$ is a K-basis of F.

Proof: If K is a subfield of E and E is a subfield of F, then certainly K is a subfield of F. Thus F is an extension of K.

Now the claim about the degree. Put $|F:E| = r$ and $|E:K| = s$ for brevity. We are to prove that the dimension of F over K is equal to rs. Let $\{f_1, f_2, \ldots f_r\}$ be an E-basis of F and $\{e_1, e_2, \ldots, e_r\}$ a K-basis of E. We are to find a K-basis of F having exactly rs elements. The most natural thing to do is to consider the rs products $f_i e_j$. We contend that $\{f_i e_j : i = 1,2,\ldots,r;\ j = 1,2,\ldots,s\}$ is a K-basis of F.

First we show that $\{f_i e_j\}$ spans F over K. Indeed, let f be an arbitrary element of F. Then

$$f = b_1 f_1 + b_2 f_2 + \cdots + b_r f_r$$

for some $b_1, b_2, \ldots, b_r \in E$, because $\{f_i : i = 1,2,\ldots,r\}$ spans F over E; and for each i,

$$b_i = a_{i1} e_1 + a_{i2} e_2 + \cdots + a_{is} e_s$$

for some $a_{i1}, a_{i2}, \ldots, a_{is} \in K$, because $\{e_j : j = 1,2,\ldots,s\}$ spans E over K. Hence

$$f = \sum_{i=1}^{r} b_i f_i = \sum_{i=1}^{r} \Big(\sum_{j=1}^{s} a_{ij} e_j \Big) f_i = \sum_{i,j} a_{ij} (e_j f_i)$$

is a linear combination of $e_j f_i = f_i e_j$ over K. Thus $\{f_i e_j\}$ spans F over K.

Furthermore, $\{f_i e_j\}$ is linearly independent over K. Indeed, if b_{ij} are elements of K such that

$$\sum_{i,j} b_{ij} f_i e_j = 0 \quad ,$$

then

$$\sum_{i=1}^{r} \Big(\sum_{j=1}^{s} b_{ij} e_j \Big) f_i = 0,$$

where $\sum_{j=1}^{s} b_{ij} e_j \in E$ for each i. Since $\{f_i : i = 1,2,\ldots,r\}$ is linearly independent over E, we have $\sum_{j=1}^{s} b_{ij} e_j = 0$ for each i. Since $\{e_j : j = 1,2,\ldots,s\}$ is linearly independent over E, we obtain $b_{ij} = 0$ for each i,j. Hence $\{f_i e_j\}$ is linearly independent over K.

Thus $\{f_i e_j\}$ is is a K-basis of F and $|F:K| = rs = |F:E|\,|E:K|$. □

It follows from Theorem 48.13 by induction that

$$|K_n:K_1| = |K_n:K_{n-1}|\,|K_{n-1}:K_{n-2}|\ldots|K_2:K_1|$$

whenever $K_n/K_{n-1}, K_{n-1}/K_{n-2}, \ldots, K_2/K_1$ are finite dimensional field extensions. In fact, Theorem 48.13 and its generalization is true for infinite dimensional extensions, too, but we will not need this.

48.14 Lemma: *Let F/E and E/K be field extensions. If $|F:K|$ is finite, then $|F:E|$ and $|E:K|$ are both finite. In fact, both of them are divisors of $|F:K|$ and $|F:K| = |F:E|\,|E:K|$.*

Proof: Let $n = |F:K|$ and let $\{f_i : i = 1,2,\ldots,n\}$ be a basis of F over K. Then $\{f_i : i = 1,2,\ldots,n\}$ spans F over E and so $|F:E| \leqslant n$ by Steinitz' replacement theorem. Thus $|F:E|$ is finite.

Now the finiteness of $|E:K|$. If E were infinite dimensional over K, there would be $n+1$ K-linearly independent elements of E, so there would be $n+1$ K-linearly independent elements of F, contradicting $|F:K| = n$. Thus $|E:K|$ is finite.

We now obtain $n = |F:K| = |F:E|\,|E:K|$ from Theorem 48.13. In particular, $|F:E|$ and $|E:K|$ divide n. □

Exercises

1. Let E be a field and $K \subseteq E$. Show that K is a subfield of E if and only if K is a subgroup of E and $K\backslash\{0\}$ is a subgroup of $E\backslash\{0\}$.

2. Let p be prime. Is $\mathbb{Z}_{p^2}$ an extension of $\mathbb{Z}_p$? Is $\mathbb{Z}_{p^3}$ an extension of $\mathbb{Z}_{p^2}$?

3. Prove that $\mathbb{Q}(\omega) = \{x + y\omega : x,y \in \mathbb{Q}\}$ and $\mathbb{Q}(\sqrt{5}i) = \{x + y\sqrt{5}i : x,y \in \mathbb{Q}\}$ are subfields of $\mathbb{C}$.

4. Let K be a field and let $Aut(K)$ be the set of all field automorphisms of K. Show that $Aut(K)$ is a group under composition.

5. Find all automorphisms of $\mathbb{Q}$, $\mathbb{F}_p$, $\mathbb{Q}(i)$, $\mathbb{Q}(\omega)$, $\mathbb{Q}(\sqrt{5}i)$, $\mathbb{Q}(\sqrt[3]{2})$ (see Ex.3).

6. Find three nonisomorphic infinite fields of characteristic $p \neq 0$.

7. Find the degrees of the following extensions: $\mathbb{C}/\mathbb{R}$, $\mathbb{C}/\mathbb{Q}(i)$, $\mathbb{Q}(i)/\mathbb{Q}$, $\mathbb{R}/\mathbb{Q}$, $\mathbb{F}(x)/\mathbb{F}$.

8. Show that $\mathbb{Q}(\sqrt{2},i) := \{a + b\sqrt{2} + ci + d\sqrt{2}i : a,b,c,d \in \mathbb{Q}\}$ is an extension field of both $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt{2})$. Find $|\mathbb{Q}(\sqrt{2},i):\mathbb{Q}|$ by two different methods.

9. Prove or disprove: If E/K_1 and E/K_2 are finite dimensional field extensions, then $E/(K_1 \cap K_2)$ is finite dimensional, too.

10. Let K be a field and e the identity element of K. Show that $char\ K = 0$ or p according as the subring of K generated by e is isomorphic to $\mathbb{Z}$ or to $\mathbb{Z}_p$.

11. Find the prime subfields of the fields in §29, Ex. 8.

12. Let K be a field of characteristic $p \neq 0$. Prove that $\varphi: K \to K$ is a field homomorphism.
$$a \to a^p$$

§49
Field Extensions (continued)

49.1 Definition: Let E be an extension field of K. If F is a field such that $K \subseteq F \subseteq E$, then F is said to be an *intermediate field of the extension E/K*.

49.2 Definition: Let E/K be a field extension and let S be a subset of E. The intersection of all subfields of E containing $K \cup S$, which is a subfield of E by Example 48.3(f), is called the *subfield of E generated by S over K*, and is denoted by $K(S)$.

It follows immediately from this definition that $K \subseteq K(S) \subseteq E$ so that $K(S)$ is an intermediate field of E/K. When S is a finite subset of E, say when $S = \{a_1, a_2, \ldots, a_n\}$, we write $K(a_1, a_2, \ldots, a_n)$ instead of $K(\{a_1, a_2, \ldots, a_n\})$. In particular, if $a \in E$, then $K(a)$ is, by definition, the smallest subfield of E containing both K and a. Notice that $K(a_1, a_2, \ldots, a_n) = K(a_{i_1}, a_{i_2}, \ldots, a_{i_n})$ for any permutation $\begin{pmatrix} 1 & 2 & \ldots & n \\ i_1 & i_2 & \ldots & i_n \end{pmatrix}$ in S_n.

49.3 Definition: Let E/K be a field extension and let S be a subset of E. The intersection of all subrings of E containing $K \cup S$, which is a subring of E by Example 30.2 (c), is called the *subring of E generated by S over K*, and is denoted by $K[S]$.

Since every subfield of E containing $K \cup S$ is also a subring of E containing $K \cup S$, we clearly have $K \subseteq K[S] \subseteq K(S) \subseteq E$. If S is a finite subset of E, say $S = \{a_1, a_2, \ldots, a_n\}$, we write $K[a_1, a_2, \ldots, a_n]$ instead of $K[\{a_1, a_2, \ldots, a_n\}]$. In particular, if $a \in E$, then $K[a]$ is, by definition, the smallest subring of E containing both K and a. We have $K[a_1, a_2, \ldots, a_n] = K[a_{i_1}, a_{i_2}, \ldots, a_{i_n}]$ for any permutation $\begin{pmatrix} 1 & 2 & \ldots & n \\ i_1 & i_2 & \ldots & i_n \end{pmatrix}$ in S_n.

49.4 Example: In the extension $\mathbb{C}/\mathbb{Q}$, let us find the subfield of $\mathbb{C}$ generated by i over $\mathbb{Q}$. Any subfield of $\mathbb{C}$ containing both $\mathbb{Q}$ and i contains complex numbers of the form $\frac{a+bi}{c+di}$, where $a,b,c,d \in \mathbb{Q}$ and $c+di \neq 0$. One verifies easily that $F = \left\{ \frac{a+bi}{c+di} \in \mathbb{C} : a,b,c,d \in \mathbb{Q}, c+di \neq 0 \right\}$ is a subfield of $\mathbb{C}$ containing both $\mathbb{Q}$ and i. Hence F is the subfield of $\mathbb{C}$ generated by i over $\mathbb{C}$.

Let us note that any element of F can be written in the form $x+yi$, with $x,y \in \mathbb{Q}$. Thus $\{x+yi \in \mathbb{C} : x,y \in \mathbb{Q}\} = F$ and F is equal to the field $\mathbb{Q}(i)$ defined in Example 48.3(c). So the notation of Example 48.3(c) is consistent with that of Definition 49.2.

The description of the elements in a field generated by a subset over a subfield resembles the preceding example.

49.5 Lemma: *Let E/K be a field extension and $a_1,a_2,\ldots,a_n \in E$. Then*
(1) $K[a_1,a_2,\ldots,a_n] = \{f(a_1,a_2,\ldots,a_n) \in E : f \in K[x_1,x_2,\ldots,x_n]\}$;
(2) $K(a_1,a_2,\ldots,a_n)$

$$= \left\{ \frac{f(a_1,a_2,\ldots,a_n)}{g(a_1,a_2,\ldots,a_n)} \in E : f,g \in K[x_1,x_2,\ldots,x_n],\ g(a_1,a_2,\ldots,a_n) \neq 0 \right\}.$$

Proof: (1) Let A be the set on the right hand side of the equation in (1). Any subring of E containing K and $\{a_1,a_2,\ldots,a_n\}$ will contain the elements of the form $k a_1^{m_1} a_2^{m_2} \ldots a_n^{m_n}$, where $k \in K$ and $m_1,m_2,\ldots,m_n$ are nonnegative integers, hence also the elements of the form

$$\sum k_{m_1 m_2 \ldots m_n} a_1^{m_1} a_2^{m_2} \ldots a_n^{m_n} \qquad (*)$$

where $k_{m_1 m_2 \ldots m_n} \in K$ and $m_1,m_2,\ldots,m_n$ are nonnegative integers. Of course, (*) is nothing but the value of the polynomial

$$f(x_1,x_2,\ldots,x_n) = \sum k_{m_1 m_2 \ldots m_n} x_1^{m_1} x_2^{m_2} \ldots x_n^{m_n} \in K[x_1,x_2,\ldots,x_n]$$

at $(a_1,a_2,\ldots,a_n)$. So every element of A is in any subring of E containing K and $\{a_1,a_2,\ldots,a_n\}$. This gives $A \subseteq K[a_1,a_2,\ldots,a_n]$. To prove the reverse inclusion, it suffices, in view of $K \cup \{a_1,a_2,\ldots,a_n\} \subseteq A \subseteq E$, to show that A is a subring of E. But this is immediate: given any $f(a_1,a_2,\ldots,a_n)$ and $g(a_1,a_2,\ldots,a_n) \in A$, where $f,g \in K[x_1,x_2,\ldots,x_n]$, we have

$$\begin{aligned} f(a_1,a_2,\ldots,a_n) + g(a_1,a_2,\ldots,a_n) &= (f+g)(a_1,a_2,\ldots,a_n) \quad \in A \\ -g(a_1,a_2,\ldots,a_n) &= (-g)(a_1,a_2,\ldots,a_n) \quad \in A \\ f(a_1,a_2,\ldots,a_n)g(a_1,a_2,\ldots,a_n) &= (fg)(a_1,a_2,\ldots,a_n) \quad \in A \end{aligned}$$

since $f+g$, $-g$, fg belong to $[x_1,x_2,\ldots,x_n]$ whenever f,g do. Thus A is a subring of E by the subring criterion (Lemma 30.2). This proves $K[a_1,a_2,\ldots,a_n] = A$.

(2) The reasoning is similar. Let B be the set on the right hand side of the equation in (2). Clearly $A \subseteq B$. Note that $B = \{b/c \in E: b,c \in A, c \neq 0\} = \{bc^{-1} \in E: b,c \in A, c \neq 0\}$. Any subfield of E containing K and $\{a_1,a_2,\ldots,a_n\}$ will contain $K[a_1,a_2,\ldots,a_n] = A$ and, since a subfield is closed under division, it will contain also the elements b/c, where $b,c \in A$ and $c \neq 0$. This means that B is contained in any subfield of E containing K and $\{a_1,a_2,\ldots,a_n\}$. Hence $B \subseteq K(a_1,a_2,\ldots,a_n)$. To prove the reverse inclusion, it suffices, in view of $K \cup \{a_1,a_2,\ldots,a_n\} \subseteq B \subseteq E$, to show that B is a subfield of E. Indeed, given any b/c, $d/e \in B$, where $b,c,d,e \in A$, $c,e \neq 0$, we have

$$\begin{aligned} \frac{b}{c} + \frac{d}{e} &= \frac{be+dc}{ce} \in B \\ -\frac{d}{e} &\in B \\ \frac{b}{c}\frac{d}{e} &= \frac{bd}{ce} \in B \\ \frac{1}{\frac{d}{e}} &= \frac{e}{d} \in B \quad \text{(provided } d/e \neq 0, \text{ i.e., } d \neq 0) \end{aligned}$$

since $be+dc$, ce, $-d$, bd, ce belong to A whenever b,c,d,e do and $ce \neq 0$ whenever $c \neq 0 \neq e$ (A is a subring of the field E and has therefore no zero divisors). Thus B is a subfield of E by the subfield criterion (Lemma 48.2). This proves $K(a_1,a_2,\ldots,a_n) = B$. □

The proof of Lemma 49.5 can be somewhat simplified by referring to Theorem 31.8.

Let us take a new look at Example 49.4 under the light of Lemma 49.5. The field F in Example 49.4 is exactly the the field desribed in Lemma 49.5, with $K = \mathbb{Q}$, $n = 1$, $a_1 = i \in \mathbb{C}$. On the other hand, the field $\{x + yi \in \mathbb{C} : x,y \in \mathbb{Q}\}$ is exactly the subring of $\mathbb{C}$ described in Lemma 49.5, with $K = \mathbb{Q}$, $n = 1$, $a_1 = i \in \mathbb{C}$. Thus we have $\mathbb{Q}(i) = \mathbb{Q}[i]$. The reader will easily verify that $\mathbb{Q}(\sqrt{2}) = \mathbb{Q}[\sqrt{2}]$ also (cf. Theorem 50.6).

49.6 Lemma: *Let E/K be a field extension and $a,b,a_1,a_2,\ldots,a_n \in E$.*

(1) *$K(a) = K$ if and only if $a \in K$.*

(2) *$K(a_1,a_2,\ldots,a_{n-1},a_n) = (K(a_1,a_2,\ldots,a_{n-1}))(a_n)$ and $K[a_1,a_2,\ldots,a_{n-1},a_n] = [K(a_1,a_2,\ldots,a_{n-1}]][a_n]$.*

(3) *$K(a,b) = (K(a))(b) = (K(b))(a)$ and $K[a,b] = [K[a]][b] = [K[b]][a]$.*

Proof: (1) $a \in K(a)$ by the definition of $K(a)$ and, if $K(a) = K$, we obtain $a \in K$. Conversely, if $a \in K$, then $K = K \cup \{a\}$ and K is the intersection of all subfields of E containing both K and a; thus $K(a) = K$.

(2) Let us write $L = K(a_1,a_2,\ldots,a_{n-1})$. Then L contains K and $a_1,a_2,\ldots,a_{n-1}$. Now $L(a_n)$ is a subfield of E containing both L and a_n, so $L(a_n)$ is a subfield of E containing K and $a_1,a_2,\ldots,a_{n-1}$ and a_n. Then $K(a_1,a_2,\ldots,a_{n-1},a_n)$, being the intersection of all subfield of E containing K and $a_1,a_2,\ldots,a_{n-1},a_n$, is a subfield of $L(a_n)$. This gives $K(a_1,a_2,\ldots,a_{n-1},a_n) \subseteq L(a_n)$. On the other hand, $K(a_1,a_2,\ldots,a_{n-1},a_n)$ is a subfield of E containing K, $a_1,a_2,\ldots,a_{n-1}$ and also a_n. So $L \subseteq K(a_1,a_2,\ldots,a_{n-1},a_n)$ by the definition of $L = K(a_1,a_2,\ldots,a_{n-1})$; and $a_n \in K(a_1,a_2,\ldots,a_{n-1},a_n)$. Hence $K(a_1,a_2,\ldots,a_{n-1},a_n)$ is a subfield of E containing both L and a_n. Then $L(a_n) \subseteq K(a_1,a_2,\ldots,a_{n-1},a_n)$ by the definition of $L(a_n)$. We obtain $K(a_1,a_2,\ldots,a_{n-1},a_n) = L(a_n)$, as was to be proved. The second assertion is proved in exactly the same way (read "subring" in place of "subfield" in the foregoing argument).

(3) Using part (2) twice, we get $(K(a))(b) = K(a,b) = K(b,a) = (K(b))(a)$ and similarly $[K[a]][b] = K[a,b] = K[b,a] = [K[b]][a]$. □

We introduce a very important classification of field extensions: algebraic vs. transcendental extensions. They behave very differently.

49.7 Definition: Let E/K be a field extension. An element a of E is said to be *algebraic over* K if there is a nonzero polynomial f in $K[x]$ such that a is a root of f, i.e., $f(a) = 0$. An element a of E is said to be *transcendental over* K if a is not algebraic, that is to say, if there is no nonzero polynomial f in $K[x]$ with $f(a) = 0$.

If every element of E is algebraic over K, then E is called an *algebraic extension of* K and E/K is called an *algebraic extension*. In this case, E is said to be *algebraic over* K. If E is not an algebraic extension of K, then E is called a *transcendental extension of* K and E/K is called a *transcendental extension*. If so, that is to say, if E contains at least one element which is not algebraic over K, then E is said to be *transcendental over* K.

49.8 Examples: (a) Let K be any field. Then, for any element $a \in K$, the polynomial $f_a(x) := x - a$ is in $K[x]$, and a is a root of f_a. Thus any element of K is algebraic over K, and K is an algebraic extension of K.

(b) $i \in \mathbb{C}$ is a root of the polynomial $x^2 + 1 \in \mathbb{Q}[x]$. Hence i is algebraic over $\mathbb{Q}$. Also, any element $a + bi$ of $\mathbb{Q}(i)$, where $a,b \in \mathbb{Q}$, is a root of

$$[x - (a + bi)][x - (a - bi)] = x^2 - 2ax + (a^2 + b^2) \in \mathbb{Q}[x]$$

and is therefore algebraic over $\mathbb{Q}$. Hence $\mathbb{Q}(i)/\mathbb{Q}$ is an algebraic extension.

(c) $\sqrt{2} \in \mathbb{R}$ is a root of the polynomial $x^2 - 2 \in \mathbb{Q}[x]$. Hence $\sqrt{2}$ is algebraic over $\mathbb{Q}$. Also, any element $a + b\sqrt{2}$ of $\mathbb{Q}(\sqrt{2})$, where $a,b \in \mathbb{Q}$, is a root of

$$[x - (a + b\sqrt{2})][x - (a - b\sqrt{2})] = x^2 - 2ax + (a^2 - 2b^2) \in \mathbb{Q}[x]$$

and is therefore algebraic over $\mathbb{Q}$. Hence $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ is an algebraic extension.

(d) It is a fact that $\pi \in \mathbb{R}$ and $e \in \mathbb{R}$ are transcendental over $\mathbb{Q}$. We borrow this fact from number theory without proof. Thus $\mathbb{R}/\mathbb{Q}$ is a transcendental extension. $\mathbb{Q}(\pi)$ and $\mathbb{Q}(e)$ are also transcendental extensions of $\mathbb{Q}$.

(e) Let K be a field and x an indeterminate over K. Then $K(x)$ is an extension field of K and $x \in K(x)$. If f is any nonzero polynomial in $K[x]$,

then $f(x) = f \neq 0$ (Example 35.2(d)). Thus x is transcendental over K and $K(x)/K$ is a transcendental extension.

Likewise $f(x^2) \neq 0$ for any nonzero polynomial f in $K[x]$ and x^2 is transcendental over K. On the other hand, if y is another indeterminate over K, then x is the root of the polynomial $y^2 - x^2 \in (K(x^2))[y]$, so x is algebraic over $K(x^2)$. Thus an element may be transcendental over a field and algebraic over another field.

49.9 Definition: Let E/K be a field extension. If there is an element a in E such that $E = K(a)$, then E is called a *simple extension of* K. In this case, any element a of E satisfying $E = K(a)$ is called a *primitive element of the extension* E/K. If there are finitely many elements $a_1, a_2, \ldots, a_n$ in E such that $E = K(a_1, a_2, \ldots, a_n)$, then E is said to be *finitely generated over* K.

The reader should clearly distinguish between finite dimensional extensions and finitely generated extensions.

We close this paragraph with a theorem that describes all simple transcendental extensions up to isomorphism. Simple algebraic extensions will be treated in the next paragraph.

49.10 Theorem: *Let E/K be a field extension and let $a \in E$ be transcendental over K. Then $K(a) \cong K(x)$, where x is an indeterminate over K.*

Proof: We wish to find an isomorphism from $K(x)$ onto $K(a)$. What is more natural than the extension

$$\varphi: K(x) \to K(a)$$
$$\frac{f}{g} \to \frac{f(a)}{g(a)}$$

of the substitution homomorphism? In any case, Lemma 49.5(2) suggests that we try this mapping. Now φ is meaningful, for, given any $f/g \in K(x)$ with $f, g \in K[x]$, $g \neq 0$, we have $g(a) \neq 0$ (a is transcendental over K) and so $(f/g)\varphi = f(a)/g(a)$ is a perfectly definite element of $K(a)$.

We claim that φ is well defined. Indeed, if $f/g = f_1/g_1$ in $K(x)$, where $f, g, f_1, g_1 \in K[x]$ and $g \neq 0 \neq g_1$, then $fg_1 = f_1 g$ in $K[x]$ by Lemma 35.3,

$f(a)g_1(a) = f_1(a)g(a)$ in E, with $g_1(a) \neq 0 \neq g(a)$; multiplying this equation by $1/g_1(a)g(a)$, we obtain

$$\left(\frac{f}{g}\right)\varphi = \frac{f(a)}{g(a)} = \frac{f_1(a)}{g_1(a)} = \left(\frac{f_1}{g_1}\right)\varphi,$$

which shows that φ is well defined.

φ is a ring homomorphism because, from Lemma 35.3, we have

$$\left(\frac{f}{g} + \frac{p}{q}\right)\varphi = \frac{fq + pg}{gq}\varphi = \frac{(fq + pg)(a)}{(gq)(a)} = \frac{f(a)q(a) + p(a)g(a)}{g(a)q(a)}$$
$$= \frac{f(a)}{g(a)} + \frac{p(a)}{q(a)} = \left(\frac{f}{g}\right)\varphi + \left(\frac{p}{q}\right)\varphi$$

and $\left(\frac{f}{g}\frac{p}{q}\right)\varphi = \frac{fp}{gq}\varphi = \frac{f(a)p(a)}{g(a)q(a)} = \frac{f(a)}{g(a)}\frac{p(a)}{q(a)} = \left(\frac{f}{g}\right)\varphi\left(\frac{p}{q}\right)\varphi$

for any f/g, $p/q \in K(x)$, where $f,g,p,q \in K[x]$ and $g \neq 0 \neq q$, the last condition ensuring $g(a) \neq 0 \neq q(a)$.

Since
$$\begin{aligned} Ker\,\varphi &= \{f/g \in K(x):\ f,g \in K[x],\ g \neq 0 \text{ in } K[x],\ f(a)/g(a) = 0\} \\ &= \{f/g \in K(x):\ f,g \in K[x],\ g \neq 0,\ f(a) = 0\} \\ &= \{f/g \in K(x):\ f,g \in K[x],\ g \neq 0,\ f = 0\} \\ &= \{0\}, \end{aligned}$$

φ is one-to-one. Hence φ is a field homomorphism. Lemma 49.5(2) states that φ is onto $K(a)$. So $\varphi: K(x) \to K(a)$ is a field isomorphism: $K(x) \cong K(a)$. □

Exercises

1. Let E/K be a field extension and $S \subseteq E$, $S \neq \varnothing$. Show that
$K[S] = \{f(s_1,s_2,\ldots,s_n) \in E:\ n \in \mathbb{N}, f \in K[x_1,x_2,\ldots,x_n]$ and $s_1,s_2,\ldots,s_n \in S\}$;
and $K(S)$

$$= \left\{\frac{f(s_1,s_2,\ldots,s_n)}{g(s_1,s_2,\ldots,s_n)} \in E:\ n \in \mathbb{N}, f,g \in K[x_1,x_2,\ldots,x_n] \text{ and } g(s_1,s_2,\ldots,s_n) \neq 0\right\}.$$

2. Let E/K be a field extension and $S \subseteq E$, $S \neq \varnothing$. Using the definition of $K[S]$ and $K(S)$ only (in particular, without using Ex. 1), prove that $K(S)$ is the field of fractions of $K[S]$.

3. Let E/K be a field extension and $S \subseteq E$. Show that $K(S) = K$ if and only if $S \subseteq K$.

4. Let E/K be a field extension and $a_1, a_2, \ldots, a_n \in E$. Prove that $(K(a_1, \ldots, a_k))(a_{k+1}, \ldots, a_n)$ for any $k = 1,2, \ldots, n-1$.

5. Let a,b are arbitrary rational numbers. Find a polynomial in $\mathbb{Q}[x]$ which admits $a + b\sqrt{5}$ as a root. Conclude that $\mathbb{Q}(\sqrt{5})/\mathbb{Q}$ is an algebraic extension.

6. Show that $\sqrt{2} + i$, $\sqrt{2} + \sqrt{3}$, $\sqrt{2} + \sqrt{3} + i$ are algebraic over $\mathbb{Q}$ by exhibiting polynomials in $\mathbb{Q}[x]$ having these numbers among their roots.

7. Let K be a field. Prove that every element in $K(x)\backslash K$ is transcendental over K.

8. Let E/K be a simple field extension and let a be a primitive element of this extension. Let $k,k' \in K$, with $k \neq 0$. Show that $ka + k'$ is also a primitive element of E/K.

9. Find a finitely generated field extension which is not finite dimensional. Prove that every finite dimensional extension is finitely generated.

10. Prove or disprove: if E/K is a field extension and $a,b \in E$ are transcendental over K, then $K(a,b) \cong K(x,y)$, where x,y are indeterminates over K.

§50
Algebraic Extensions

Let E/K be a field extension and let $a \in E$ be algebraic over K. Then there is a nonzero polynomial f in $K[x]$ such that $f(a) = 0$. Hence the subset $A = \{f \in K[x]: f(a) = 0\}$ of $K[x]$ does not consist only of 0. We observe that A is an ideal of $K[x]$, because A is the kernel of the substitution homomorphism $T_a: K[x] \to E$.

Thus A is an ideal of $K[x]$ and $A \neq \{0\}$. Since $K[x]$ is a principal ideal domain, $A = K[x]f_0 =: (f_0)$ for some nonzero polynomial f_0 in $K[x]$. For any polynomial $g \in K[x]$, the relation $(g) = A = (f_0)$ holds if and only if g and f_0 are associate in $K[x]$, that is to say, if and only if $g(x) = cf_0(x)$ for some c in $K^\times$. There is a unique $c_0 \in K^\times$ such that the leading coefficient of $c_0 f_0(x)$ is equal to 1. With this c_0, we put $g_0(x) = c_0 f_0(x)$. Then g_0 is the unique monic polynomial in $K[x]$ satisfying $(g_0) = A = \{f \in K[x]: f(a) = 0\}$, and $f(a) = 0$ for a polynomial f in $K[x]$ if and only if $g_0 | f$ in $K[x]$. In particular, we have $deg\ g_0 \leqslant deg\ f$ for any $f \in K[x]$ having a as a root.

In this way, we associate with $a \in E$ a unique monic polynomial g_0 in $K[x]$. This g_0 is the monic polynomial in $K[x]$ of least degree having a as a root.

g_0 is irreducible over K: if there are polynomials $p(x)$, $q(x)$ in $K[x]$ with $g_0(x) = p(x)q(x)$, $1 \leqslant deg\ p(x) < deg\ g_0(x)$ and $1 \leqslant deg\ q(x) < deg\ g_0(x)$, then $0 = g_0(a) = p(a)q(a)$ would imply $p(x) \in A$ or $q(x) \in A$, hence $g_0 | p$ or $g_0 | q$ in $K[x]$, which is impossible in view of the conditions on $deg\ p(x)$ and $deg\ q(x)$.

We proved the following theorem.

50.1 Theorem: *Let E/K be a field extension and $a \in E$. If a is algebraic over K, then there is a unique nonzero monic polynomial $g(x)$ in $K[x]$ such that*

$$\text{for all } f(x) \in K[x], \qquad f(x) = 0 \text{ if and only if } g(x) | f(x) \text{ in } K[x].$$

In particular, a is a root of $g(x)$ and $g(x)$ has the smallest degree among the nonzero polynomials in $K[x]$ admitting a as a root. Moreover, $g(x)$ is irreducible over K. □

50.2 Definition: Let E/K be a field extension and let $a \in E$ be algebraic over K. The unique polynomial $g(x)$ of Theorem 50.1 is called the *minimal polynomial of a over K.*

The minimal polynomial of a over K is also called the *irreducible polynomial of a over K.* Given an element a of E, algebraic over K, and a polynomial $h(x)$ in $K[x]$, in order to find out whether $h(x)$ is the minimal polynomial of a over K, it seems we had to check whether $h(x)|f(x)$ for all the polynomials $f(x) \in K[x]$ having a as a root. Fortunately, there is another characterization of minimal polynomials.

50.3 Theorem: *Let E/K be a field extension and $a \in E$. Assume that a is algebraic over K. Let $h(x)$ be a nonzero polynomial in $K[x]$. If*

(i) *$h(x)$ is monic,*

(ii) *a is a root of $h(x)$,*

(iii) *$h(x)$ is irreducible over K,*

then $h(x)$ is the minimal polynomial of a over K.

Proof: We must show only that $h(x)$ divides any polynomial $f(x) \in K[x]$ having a as a root. Let $f(x)$ be a polynomial in $K[x]$ and assume that a is a root of $f(x)$. We divide $f(x)$ by $h(x)$ and get

$$f(x) = q(x)h(x) + r(x), \qquad r(x) = 0 \text{ or } deg\ r(x) < deg\ h(x)$$

with suitable $q(x), r(x) \in K[x]$. Substituting a for x, we obtain

$$0 = f(a) = q(a)h(a) + r(a) = q(a)0 + r(a) = r(a).$$

If $r(x)$ were distinct from the zero polynomial in $K[x]$, then the irreducible polynomial $h(x)$ would have a common root a with the polynomial $r(x)$ whose degree is smaller than the degree of $h(x)$. This is impossible by Theorem 35.18(4). Hence $r(x) = 0$ and $f(x) = q(x)h(x)$. Therefore $h(x)|f(x)$ for any polynomial $f(x) \in K[x]$ having a as a root, as was to be proved. □

50.4 Examples: (a) Let us find the minimal polynomial of $i \in \mathbb{C}$ over $\mathbb{R}$. Since i is a root of the polynomial $x^2 + 1 \in \mathbb{R}[x]$, which is monic and irreducible over $\mathbb{R}$, Theorem 50.3 tells us that $x^2 + 1$ is the minimal polynomial of i over $\mathbb{R}$. In the same way, we see that $x^2 + 1 \in \mathbb{Q}[x]$ is the minimal polynomial of i over $\mathbb{Q}$. On the other hand, $x^2 + 1 \in (\mathbb{Q}(i))[x]$ is not irreducible over $\mathbb{Q}(i)$, because $x^2 + 1 = (x - i)(x + i)$ in $(\mathbb{Q}(i))[x]$. Now $x - i$ is a monic irreducible polynomial in $(\mathbb{Q}(i))[x]$ having i as a root, and thus $x - i$ is the minimal polynomial of $i \in \mathbb{C}$ over $\mathbb{Q}(i)$.

(b) Let us find the minimal polynomial of $u = \sqrt{2} + \sqrt{3} \in \mathbb{R}$, over $\mathbb{Q}$. The calculations

$$\begin{aligned} u &= \sqrt{2} + \sqrt{3} \\ u - \sqrt{2} &= \sqrt{3} \\ u^2 - 2\sqrt{2}u + 2 &= 3 \\ u^2 - 1 &= 2\sqrt{2}u \\ u^4 - 2u^2 + 1 &= 8u^2 \\ u^4 - 10u^2 + 1 &= 0 \end{aligned} \qquad \text{(u)}$$

show that $\sqrt{2}+\sqrt{3}$ is a root of the monic polynomial $f(x) = x^4 - 10x^2 + 1$ in $\mathbb{Q}[x]$. We will prove that $f(x)$ is irreducible over $\mathbb{Q}$. Theorem 50.3 will then yield that $f(x)$ is the minimal polynomial of $\sqrt{2}+\sqrt{3}$ over $\mathbb{Q}$.

In view of Lemma 34.11, it will be sufficient to show that $f(x)$ is irreducible over $\mathbb{Z}$. Since the numbers $\mp 1/\mp 1 = \mp 1$ are not roots of $f(x)$, we learn from Theorem 35.10 (rational root theorem) that $f(x)$ has no polynomial factor in $\mathbb{Z}[x]$ of degree one. If there were a factorization in $\mathbb{Z}[x]$ of $f(x)$ into two polynomials of degree two, which we may assume to be

$$x^4 - 10x^2 + 1 = (x^2 + ax + b)(x^2 + cx + d) \qquad \text{(e)}$$

without loss of generality, then the integers a,b,c,d would satisfy

$$a + c = 0, \quad d + ac + b = -10, \quad ad + bc = 0, \quad bd = 1$$

and this would force $b = d = \mp 1$ and the first two equations would give

$$a + c = 0, ac = -12 \quad \text{, or} \quad a + c = 0, ac = -8$$
$$a^2 = 12 \quad \text{or} \quad a^2 = 8,$$

whereas no integer has a square equal to 8 or 12. Thus $f(x)$ is irreducible in $\mathbb{Z}[x]$ and, as remarked earlier, $f(x)$ is therefore the minimal polynomial of $\sqrt{2} + \sqrt{3}$ over $\mathbb{Q}$.

The irreducibility of $f(x)$ of degree four over $\mathbb{Q}$ could be proved by showing the irreducibility of another polynomial, of degree *less* than four, over a field *larger* than $\mathbb{Q}$. As this gives a deeper insight to the

problem at hand, we will discuss this method. The equation (u) states that $\sqrt{2}+\sqrt{3}$ is a root of the polynomial $f_2(x) = x^2 - 2\sqrt{2}x - 1 \in (\mathbb{Q}(\sqrt{2}))[x]$. Let $g(x) \in (\mathbb{Q}(\sqrt{2}))[x]$ be the minimal polynomial of $\sqrt{2}+\sqrt{3}$ over $\mathbb{Q}(\sqrt{2})$. Then $g(x)|f_2(x)$ in $(\mathbb{Q}(\sqrt{2}))[x]$ and, if $g(x) \neq f_2(x)$, then *deg* $g(x)$ would be one and $g(x)$ would be $x - (\sqrt{2}+\sqrt{3})$, since the latter is the unique monic polynomial of degree one having $\sqrt{2}+\sqrt{3}$ as a root. But $g(x) \in (\mathbb{Q}(\sqrt{2}))[x]$ and this would imply $\sqrt{2}+\sqrt{3} \in \mathbb{Q}(\sqrt{2})$, so $\sqrt{3} \in \mathbb{Q}(\sqrt{2})$, so $\sqrt{3} = m + n\sqrt{2}$ with suitable $m,n \in \mathbb{Q}$, where certainly $m \neq 0 \neq n$, so $3 = m^2 + 2\sqrt{2}mn + n^2$, so $\sqrt{2} = (3 - m^2 - 2n^2)/2mn$ would be a rational number, a contradiction. Thus $f_2(x) = g(x)$ is the minimal polynomial of $\sqrt{2}+\sqrt{3}$ over $\mathbb{Q}(\sqrt{2})$.

Now the irreducibility of $f(x)$ over $\mathbb{Q}$ follows very easily. $f(x)$ has no factor of degree one in $\mathbb{Q}[x]$. If $f(x)$ had a factorization (e) in $\mathbb{Q}[x]$, where a,b,c,d are rational numbers (not necessarily integers), then $\sqrt{2}+\sqrt{3}$ would be a root of one of the factors on the right hand side of (e), say of $x^2 + ax + b$. But then $x^2 + ax + b$, being a polynomial in $(\mathbb{Q}(\sqrt{2}))[x]$ having $\sqrt{2}+\sqrt{3}$ as a root, would be divisible, in $(\mathbb{Q}(\sqrt{2}))[x]$, by the minimal polynomial $f_2(x) = x^2 - 2\sqrt{2}x - 1$ of $\sqrt{2}+\sqrt{3}$ over $\mathbb{Q}(\sqrt{2})$. Comparing degrees and leading coefficients, we would obtain $x^2 - 2\sqrt{2}x - 1 = x^2 + ax + b$, so $2\sqrt{2} = -a \in \mathbb{Q}$, a contradiction. Hence $f(x)$ is irreducible over $\mathbb{Q}$.

The next lemma crystalizes the argument employed in the last example.

50.5 Lemma: *Let $K_1 \subseteq K_2 \subseteq E$ be fields and $a \in E$. If a is algebraic over K_1, then a is algebraic over K_2. Moreover, if f_1, f_2 are, respectively, the minimal polynomials of a over K_1 and K_2, then $f_2|f_1$ in $K_2[x]$.*

Proof: If a is algebraic over K_1 and $f_1(x)$ is the minimal polynomial of a over K_1, then $f_1(a) = 0$. Since $f_1(x) \in K_1[x] \subseteq K_2[x]$, we conclude that a is algebraic over K_2. Then, from $f_1(a) = 0$ and $f_1(x) \in K_2[x]$, we obtain $f_2(x)|f_1(x)$ in $K_2[x]$ by the very definition of the minimal polynomial $f_2(x)$ of a over K_2. □

We proceed to describe simple algebraic extensions. Let us recall that we found $\mathbb{Q}[i] = \mathbb{Q}(i)$. This situation obtains whenever we consider a simple extension generated by an algebraic element.

50.6 Theorem: *Let E/K be a field extension and $a \in E$. Assume that a is algebraic over K and let f be its minimal polynomial over K. We denote by $K[x]f =: (f)$ the principal ideal generated by f in $K[x]$. Then*

$$K(a) = K[a] \cong K[x]/(f).$$

Proof: Consider the substitution homomorphism $T_a: K[x] \to E$. Here $Ker\ T_a = \{h \in K[x]: h(a) = 0\} = (f)$ by Theorem 50.1 and $Im\ T_a = K[a]$ by Lemma 49.5(1). Hence $K[x]/(f) = K[x]/Ker\ T_a \cong Im\ T_a = K[a]$.

It remains to show $K(a) = K[a]$. Since $K[a] \subseteq K(a)$, we must prove only $K(a) \subseteq K[a]$. To this end, we need only prove that $1/g(a) \in K[a]$ for any $g(x) \in K[x]$ with $g(a) \neq 0$ (Lemma 49.5). Indeed, if $g(x) \in K[x]$ and $g(a) \neq 0$, then $f \nmid g$ and, since f is irreducible in $K[x]$, the polynomials $f(x)$ and $g(x)$ are relatively prime in $K[x]$ (Theorem 35.18(3)). Thus there are polynomials $r(x)$, $s(x)$ in $K[x]$ such that

$$f(x)r(x) + g(x)s(x) = 1.$$

Substituting a for x and using $f(a) = 0$, we obtain $g(a)s(a) = 1$. Hence $1/g(a) = s(a) \in K[a]$. This proves $K[a] = K(a)$. (Another proof. Since $K[x]$ is a principal ideal domain and f is irreducible in $K[x]$, the factor ring $K[x]/(f)$ is a field by Theorem 32.25; thus $K[a]$, being a ring isomorphic to the field $K[x]/(f)$, is a subfield of E, and $K[a]$ contains K and a. So $K(a) \subseteq K[a]$ and $K(a) = K[a]$.) □

50.7 Theorem: Let *E/K be a field extension and $a \in E$. Suppose that a is algebraic over K and let f be its minimal polynomial over K. Then*

$$|K(a):K| = \deg f$$

(the degree of the field $K(a)$ over K is the degree of the minimal polynomial f in $K[x]$). *In fact, if $\deg f = n$, then $\{1, a, a^2, \ldots, a^{n-1}\}$ is a K-basis of $K(a)$ and every element in $K(a)$ can be written in the form*

$$k_0 + k_1 a + k_2 a^2 + \cdots + k_{n-1} a^{n-1} \qquad (k_0, k_1, k_2, \ldots, k_{n-1} \in K)$$

in a unique way.

Proof: We prove that $\{1,a,a^2,\ldots,a^{n-1}\}$ is a K-basis of $K(a)$. Let us show that it spans $K(a)$ over K. We know $K(a) = K[a]$ from Theorem 50.6 and $K[a] = \{g(a) \in E: g \in K[x]\}$ from Lemma 49.5(1). Thus any element u of $K(a)$ can be written as $g(a)$, where $g(x)$ is a suitable polynomial in $K[x]$. Dividing this polynomial $g(x)$ by $f(x)$, which has degree n, we get

$$g(x) = q(x)f(x) + r(x), \qquad r(x) = 0 \text{ or } deg\ r(x) \leqslant n - 1,$$

with some polynomials $q(x)$, $r(x)$ in $K[x]$. Substituting a for x, we obtain

$$u = g(a) = q(a)f(a) + r(a) = q(a)0 + r(a) = r(a).$$

If, say, $r(x) = k_0 + k_1x + k_2x^2 + \cdots + k_{n-1}x^{n-1}$, where $k_0,k_1,k_2,\ldots,k_{n-1} \in K$, then

$$u = k_0 + k_1a + k_2a^2 + \cdots + k_{n-1}a^{n-1}$$

and thus $\{1,a,a^2,\ldots,a^{n-1}\}$ spans $K(a)$ over K.

Now let us show that $\{1,a,a^2,\ldots,a^{n-1}\}$ is linearly independent over K. If $k_0,k_1,k_2,\ldots,k_{n-1}$ are elements of K such that

$$k_0 + k_1a + k_2a^2 + \cdots + k_{n-1}a^{n-1} = 0,$$

then a is a root of the polynomial $h(x) = k_0 + k_1x + k_2x^2 + \cdots + k_{n-1}x^{n-1}$ in $K[x]$, so $f(x)|h(x)$ by Theorem 50.1. Here $h(x) \neq 0$ would yield the contradiction $n = deg\ f \leqslant deg\ h \leqslant n - 1$. Therefore $h(x) = 0$, which means that $k_0 = k_1 = k_2 = \cdots = k_{n-1} = 0$. Hence $\{1,a,a^2,\ldots,a^{n-1}\}$ is linearly independent over K.

This proves $\{1,a,a^2,\ldots,a^{n-1}\}$ is a K-basis of $K(a)$. It follows that

$$|K(a):K| = dim_K K(a) = |\{1,a,a^2,\ldots,a^{n-1}\}| = n = deg\ f(x)$$

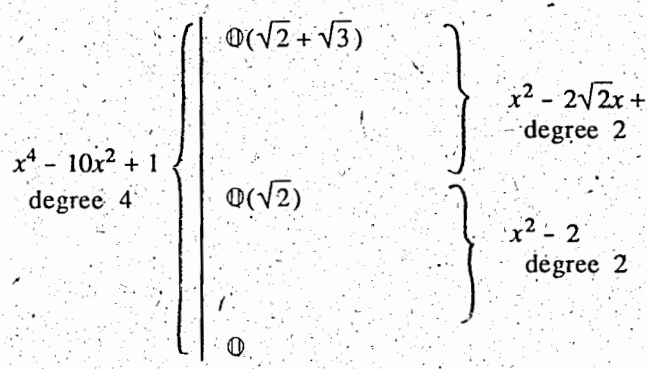
and, by Theorem 42.8, every element of $K(a)$ can be written uniquely in the form

$$k_0 + k_1a + k_2a^2 + \cdots + k_{n-1}a^{n-1}. \qquad \square$$

50.8 Definition: Let E/K be a field extension and $a \in E$. Suppose a is algebraic over K. Then the degree of its minimal polynomial over K, which is also the degree of $K(a)$ over K, is called the *degree of a over K*.

50.9 Examples: (a) The minimal polynomial of $i \in \mathbb{C}$ over $\mathbb{Q}$ is the polynomial $x^2 + 1$ in $\mathbb{Q}[x]$ (Example 50.4(a)), and $x^2 + 1$ has degree 2. Thus $i \in \mathbb{C}$ is (algebraic and) has degree 2 over $\mathbb{Q}$. Likewise, the minimal polynomial of $i \in \mathbb{C}$ over $\mathbb{R}$ is $x^2 + 1 \in \mathbb{R}[x]$ and i has degree 2 over $\mathbb{R}$.

(b) The minimal polynomial of $\sqrt{2}+\sqrt{3} \in \mathbb{R}$ over $\mathbb{Q}$ was found to be $x^4 - 10x^2 + 1 \in \mathbb{Q}[x]$ (Example 50.4(b)). Thus $\sqrt{2}+\sqrt{3}$ has degree 4 over $\mathbb{Q}$. This follows also from Theorem 50.7. In fact, the numbers 1, $\sqrt{2}$ form a $\mathbb{Q}$-basis of the field $\mathbb{Q}(\sqrt{2})$, hence $|\mathbb{Q}(\sqrt{2}):\mathbb{Q}| = 2$. Observe that



$\sqrt{2} = -\frac{9}{2}(\sqrt{2}+\sqrt{3}) + \frac{1}{2}(\sqrt{2}+\sqrt{3})^3$, so $\sqrt{2} \in \mathbb{Q}(\sqrt{2}+\sqrt{3})$ and therefore $\mathbb{Q}(\sqrt{2}) \subseteq \mathbb{Q}(\sqrt{2}+\sqrt{3})$. Thus $\mathbb{Q}(\sqrt{2})$ is an intermediate field of the extension $\mathbb{Q}(\sqrt{2}+\sqrt{3})/\mathbb{Q}$. From Theorem 48.13, we infer that

$$4 = |\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}| = |\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}(\sqrt{2})|\,|\mathbb{Q}(\sqrt{2}):\mathbb{Q}| = |\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}(\sqrt{2})|2$$

$$|\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}(\sqrt{2})| = 2$$

and $\sqrt{2}+\sqrt{3}$ has degree 2 over $\mathbb{Q}(\sqrt{2})$.

(c) Since $x^2 + 1 \in \mathbb{R}[x]$ is the minimal polynomial of $i \in \mathbb{C}$ over $\mathbb{R}$, Theorem 50.6 states that $\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{R}(i)$. In the ring $\mathbb{R}[x]/(x^2 + 1)$, we have the equality $x^2 + \mathbb{R}[x](x^2 + 1) = -1 + \mathbb{R}[x](x^2 + 1)$, and calculations are carried out just as in the ring $\mathbb{R}[x]$, but we replace $[x + \mathbb{R}[x](x^2 + 1)]^2 = x^2 + \mathbb{R}[x](x^2 + 1)$ by $-1 + \mathbb{R}[x](x^2 + 1)$. In the same way, calculations are carried out in $\mathbb{R}(i) = \mathbb{C}$ just as though i were an indeterminate over $\mathbb{R}$, and we write -1 for i^2 wherever we see i^2. This is what the isomorphism $\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{R}(i) = \mathbb{C}$ means.

(d) Likewise, if E/K is a field extension and $a \in E$, and if a is algebraic over K with the minimal polynomial $x^n + c_{n-1}x^{n-1} + c_{n-2}x^{n-2} + \cdots + c_1x + c_0$ over K so that

$$a^n = -c_{n-1}a^{n-1} - c_{n-2}a^{n-2} - \cdots - c_1a - c_0,$$

then $K(a)$ consists of the elements

$$k_0 + k_1a + \cdots + k_{n-2}a^{n-2} + k_{n-1}a^{n-1} \qquad (k_0, k_1, \ldots, k_{n-2}, k_{n-1} \in K)$$

and computations are carried out in $K(a)$ just as though a were an indeterminate over K and then replacing a^n by $-c_{n-1}a^{n-1} - c_{n-2}a^{n-2} - \cdots - c_1a - c_0$ wherever it occurs.

For instance, writing a for $\sqrt{2}+\sqrt{3} \in \mathbb{R}$, we have $a^4 = 10a^2 - 1$ in $\mathbb{Q}(a)$. If $t = 2 + a - a^2 + 3a^3 \in \mathbb{Q}(a)$ and $u = a + a^2 + 2a^3 \in \mathbb{Q}(a)$, then

$$t + u = 2 + 2a + 5a^3 \in \mathbb{Q}(a)$$

and
$$\begin{aligned} tu &= (2 + a - a^2 + 3a^3)(a + a^2 + 2a^3) \\ &= 2a + 2a^2 + 4a^3 + a^2 + a^3 + 2a^4 - a^3 - a^4 - 2a^5 + 3a^4 + 3a^5 + 6a^6 \\ &= 2a + 3a^2 + 4a^3 + 4a^4 + a^5 + 6a^6 \\ &= 2a + 3a^2 + 4a^3 + 4(10a^2 - 1) + a(10a^2 - 1) + 6a^2(10a^2 - 1) \\ &= 2a + 3a^2 + 4a^3 + 40a^2 - 4 + 10a^3 - a + 60(10a^2 - 1) - 6a^2 \\ &= -64 + a + 637a^2 + 14a^3 \in \mathbb{Q}(a). \end{aligned}$$

Let us find the inverse of $a^2 + a + 1$. According to Theorem 50.6, we must find polynomials $r(x)$, $s(x)$ in $\mathbb{Q}[x]$ such that

$$(x^4 - 10x^2 + 1)r(x) + (x^2 + x + 1)s(x) = 1$$

and this we do by the Euclidean algorithm:

$$\begin{aligned} x^4 - 10x^2 + 1 &= (x^2 - x - 10)(x^2 + x + 1) + (11x + 11) \\ x^2 + x + 1 &= (\tfrac{1}{11}x)(11x + 11) + 1, \end{aligned}$$

so that
$$\begin{aligned} 1 &= (x^2 + x + 1) - (\tfrac{1}{11}x)(11x + 11) \\ &= (x^2 + x + 1) - (\tfrac{1}{11}x)[(x^4 - 10x^2 + 1) - (x^2 - x - 10)(x^2 + x + 1)] \\ &= (x^2 + x + 1)(1 + (\tfrac{1}{11}x)(x^2 - x - 10)) - (\tfrac{1}{11}x)(x^4 - 10x^2 + 1), \\ 1 &= (x^2 + x + 1)(\tfrac{1}{11}x^3 - \tfrac{1}{11}x^2 - \tfrac{10}{11}x + 1) - (\tfrac{1}{11}x)(x^4 - 10x^2 + 1) \end{aligned}$$

and, substituting a for x, we get

$$\begin{aligned} 1 &= (a^2 + a + 1)(\tfrac{1}{11}a^3 - \tfrac{1}{11}a^2 - \tfrac{10}{11}a + 1), \\ 1/(a^2 + a + 1) &= \tfrac{1}{11}a^3 - \tfrac{1}{11}a^2 - \tfrac{10}{11}a + 1. \end{aligned}$$

Notice that a is treated here merely as a symbol that satisfies the relation $a^4 - 10a^2 + 1 = 0$. The *numerical* value of $a = \sqrt{2} + \sqrt{3} = 3.14626337\ldots$ as a real number is totaly ignored. This is algebra, the calculus of symbols. This allows enormous flexibility: we can regard a as an element in *any* extension field E of $\mathbb{Q}$ in which the polynomial $x^4 - 10x^2 + 1$ has a root. This idea will be pursued in the next paragraph.

50.10 Theorem: *Let E/K be a finite dimensional extension. Then E is algebraic over K and also finitely generated over K.*

Proof: Let $|E:K| = n \in \mathbb{N}$. To prove that E is algebraic over K, we must show that every element of a is a root of a nonzero polynomial in $K[x]$. If u is an arbitrary element of E, then the $n + 1$ elements $1,u,u^2,\ldots,u^{n-1},u^n$ of E cannot be linearly independent over K, by Steinitz' replacement theorem. Thus there are $k_0,k_1,k_2,\ldots,k_{n-1},k_n$ in K, not all of them zero, with

$$k_0 + k_1u + k_2u^2 + \cdots + k_{n-1}u^{n-1} + k_nu^n = 0.$$

Then $g(x) = k_0 + k_1x + k_2x^2 + \cdots + k_{n-1}x^{n-1} + k_nx^n$ is a nonzero polynomial in $K[x]$, in fact of degree $\leqslant n$, and u is a root of $g(x)$. Thus u is algebraic over K. Since u was arbitrary, E is algebraic over K.

Secondly, if $\{b_1,b_2,\ldots,b_n\} \subseteq E$ is a K-basis of E, then

$$\begin{aligned} E = s_K(b_1,b_2,\ldots,b_n) &= \{k_1b_1 + k_2b_2 + \cdots + k_nb_n\} \\ &\subseteq \{f(b_1,b_2,\ldots,b_n) \in E : f \in K[x_1,x_2,\ldots,x_n]\} \\ &= K(b_1,b_2,\ldots,b_n) \\ &\subseteq E, \end{aligned}$$

thus $E = K(b_1,b_2,\ldots,b_n)$ is finitely generated over K. □

As a separate lemma, we record the fact that the polynomial $g(x)$ in the preceding proof has degree $\leqslant n$.

50.11 Lemma: *Let E/K be a field extension of degree $|E:K| = n \in \mathbb{N}$. Then every element of E is algebraic over K and has degree over K at most equal to n.* □

Next we show that an extension generated by algebraic elements is algebraic.

50.12 Theorem: *Let E/K be a field extension and let $a_1,a_2,\ldots,a_{n-1},a_n$ be finitely many elements in E. Suppose that $a_1,a_2,\ldots,a_{n-1},a_n$ are algebraic over K. Then $K(a_1,a_2,\ldots,a_{n-1},a_n)$ is an algebraic extension of K. In fact, $K(a_1,a_2,\ldots,a_{n-1},a_n)$ is a finite dimensional extension of K and*

$$|K(a_1,a_2,\ldots,a_{n-1},a_n):K| \leq |K(a_1):K|\,|K(a_2):K|\ldots|K(a_n):K|$$

Proof: Let $r_1 = |K(a_1):K|$. For each $i = 2,\ldots,n-1,n$, the element a_i is algebraic over K, hence also algebraic over $K(a_1,\ldots,a_{i-1})$ by Lemma 50.5. This lemma yields, in addition, that the minimal polynomial of a_i over the field $K(a_1,\ldots,a_{i-1})$ is a divisor of the minimal polynomial of a_i over K; so, comparing the degrees of these minimal polynomials and using Theorem 50.7, we get $r_i := |(K(a_1,\ldots,a_{i-1}))(a_i):K(a_1,\ldots,a_{i-1})| \leq |K(a_i):K|$, this for all $i = 2,\ldots,n-1,n$. From

$$K \subseteq K(a_1) \subseteq K(a_1,a_2) \subseteq \ldots \subseteq K(a_1,a_2,\ldots,a_{n-1}) \subseteq K(a_1,a_2,\ldots,a_{n-1},a_n)$$

and $$K(a_1,\ldots,a_{i-1},a_i) = (K(a_1,\ldots,a_{i-1}))(a_i) \qquad \text{for } i = 2,\ldots,n-1,n$$

(Lemma 49.6(2)), we obtain

$$|K(a_1,a_2,\ldots,a_{n-1},a_n):K| = r_n r_{n-1}\ldots r_2 r_1 \qquad \text{(Theorem 48.13)}$$

$$\leq |K(a_n):K|\,|K(a_{n-1}):K|\ldots|K(a_2):K|\,|K(a_1):K|.$$

Thus $K(a_1,a_2,\ldots,a_{n-1},a_n)$ is a finite dimensional extension of K and, by Theorem 50.10, an algebraic extension of K. □

50.13 Lemma: *Let E/K be a field extension and $a,b \in E$. If a and b are algebraic over K, then $a+b$, $a-b$, ab and a/b (in case $b \neq 0$) are algebraic over K.*

Proof: If a and b are algebraic over K, then $K(a,b)$ is an algebraic extension of K by Theorem 50.12: every element of $K(a,b)$ is algebraic over K. Since $a+b$, $a-b$, ab and a/b are in $K(a,b)$, they are algebraic over K. □

50.14 Theorem: *Let E/K be a field extension and let A be the set of all elements of E which are algebraic over K. Then A is a subfield of E (and an intermediate field of the extension E/K).*

Proof: If $a,b \in A$, then a and b are algebraic over K, then $a+b$, $-b$, ab and $1/b$ (the last in case $b \neq 0$) are algebraic over K by Lemma 50.13 and so A is a subfield of E by Lemma 48.2. Since any element of K is algebraic over K (Example 49.8(a)), we have $K \subseteq A$. Thus A is an intermediate field of E/K. □

50.15 Definition: Let E/K be a field extension and let A be the subfield of E in Theorem 50.14 consisting exactly of the elements of E which are algebraic over K. Then A is called the *algebraic closure of K in E*.

A is of course an algebraic extension of K. In fact, if $a \in E$, then a is algebraic over K if and only if $a \in A$; and if F is an intermediate field of E/K, then F is algebraic over K if and only if $F \subseteq A$.

The last theorem in this paragraph states that an algebraic extension of an algebraic extension is an algebraic extension, sometimes referred to as the transitivity of algebraic extensions.

50.16 Theorem: *Let F,E,K be fields. If F is an algebraic extension of E and E is an algebraic extension of K, then F is an algebraic extension of K.*

Proof: We must show that every element of F is algebraic over K. Let $u \in F$. Since F is algebraic over E, its element u is algebraic over E, and there is a polynomial $f(x) \in E[x]$ with $f(u) = 0$, say

$$f(x) = e_0 + e_1x + \cdots + e_nx^n.$$

We put $L = K(e_0,e_1, \ldots,e_n)$. Then clearly $f(x) \in L[x]$. Since E is algebraic over K, each of $e_0,e_1, \ldots,e_n$ is algebraic over K and Theorem 50.12 tells us that L/K is finite dimensional. Also, since $f(u) = 0$ and $f(x) \in L[x]$, we see that u is algebraic over L and Theorem 50.7 tells us that $L(u)/L$ is finite dimensional. So $|L(u):K| = |L(u):L|\,|K(e_0,e_1, \ldots,e_n):K|$ is a finite number: $L(u)$ is a finite dimensional extension of K. By Theorem 50.10, $L(u)$ is an alge-

braic extension of K. So every element of $L(u)$ is algebraic over K. In particular, since $u \in L(u)$, we see that u is algebraic over K. Since u is an arbitrary element of F, we conclude that F is an algebraic extension of K. □

50.17 Definition: Let K and L be subfields of a field E. The subfield of E generated by $K \cup L$ over P, where P is the prime subfield of E, is called the *compositum of K and L*, and denoted by KL.

So $KL = P(K \cup L)$ by definition. It follows immediately from this definition that $KL = LK$. The compositum KL is the smallest subfield of E containing both K and L, whence $KL = K(L) = L(K)$.

In order to define the compositum of two fields K and L, it is necessary that these be contained in a larger field. If K and L are not subfields of a common field, we cannot define the compositum KL.

If E/K is a field extension and $a,b \in E$, then the compositum $K(a)K(b)$ of $K(a)$ and $K(b)$ is $K(P \cup \{a,b\}) = K(a,b)$.

Exercises

1. Find the minimal polynomials of the following numbers over the fields indicated.

(a) $\sqrt{2}$ over $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt{3})$.

(b) $\sqrt{3} - \sqrt{2}$ over $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt{3})$.

(c) $\sqrt{2} + \sqrt{3} + \sqrt{5}$ over $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt{3})$, $\mathbb{Q}(\sqrt{2} + \sqrt{5})$.

(d) $\sqrt[3]{2} + \sqrt{2}$ over $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt[3]{2})$, $\mathbb{Q}(\sqrt[4]{2})$.

(e) $\sqrt[3]{2} + \sqrt{3}$ over $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt[3]{2})$, $\mathbb{Q}(\sqrt{2} + \sqrt{3})$.

(f) $\sqrt{3 + \sqrt{2}}$ over $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$.

(g) $\sqrt[3]{-1+\sqrt{2}}$ over $\mathbb{Q}, \mathbb{Q}(\sqrt{2}), \mathbb{Q}(i)$.

(h) $\sqrt[3]{-1-\sqrt{2}}$ over $\mathbb{Q}, \mathbb{Q}(\sqrt{2}), \mathbb{Q}(i)$.

(j) $\sqrt[3]{-1+\sqrt{2}}+\sqrt[3]{-1-\sqrt{2}}$ over $\mathbb{Q}, \mathbb{Q}(\sqrt{2}), \mathbb{Q}(i)$.

2. Let E/K be an extension of fields and let D be an integral domain such that $K \subseteq D \subseteq E$. Prove that, if E is algebraic over K, then D is a field.

3. Let E/K be an extension of fields and $a_1, a_2, \ldots, a_m$ elements of E which are algebraic over K. Prove that $K[a_1, a_2, \ldots, a_m] = K(a_1, a_2, \ldots, a_m)$.

4. Let E/K be a field extension and $a, b \in E$. If a is algebraic of degree m over K and b is algebraic of degree n over K, show that $K(a,b)$ is an algebraic extension of K and that $|K(a,b):K| \leqslant mn$. If, in addition, m and n are relatively prime, then in fact $|K(a,b):K| = mn$.

5. Let E/K be a field extension and L, M intermediate fields. Prove the following statements.

(a) $|LM:K|$ is finite if and only if both $|L:K|$ and $|M:K|$ are finite.

(b) If $|LM:K|$ is finite, then $|L:K|$ and $|M:K|$ divide $|LM:K|$.

(c) If $|L:K|$ and $|M:K|$ are finite and relatively prime, then $|LM:K|$ is equal to $|L:K||M:K|$.

(d) If L and M are algebraic over K, then LM is algebraic over K.

(e) If L is algebraic over K, then LM is algebraic over M.

6. A complex number u is said to be an algebraic integer if u is the root of a monic polynomial in $\mathbb{Z}[x]$. Prove the following statements.

(a) If $c \in \mathbb{C}$ is algebraic over $\mathbb{Q}$, then there is a natural number n such that nc is an algebraic integer.

(b) If $u \in \mathbb{Q}$ and u is an algebraic integer, then $u \in \mathbb{Z}$.

(c) Let $f(x)$ and $g(x)$ be monic polynomials in $\mathbb{Q}[x]$. If $f(x)g(x) \in \mathbb{Z}[x]$, then $f(x)$ and $g(x)$ are in $\mathbb{Z}[x]$. (Hint: consider contents.)

(d) If $u \in \mathbb{C}$ is an algebraic integer, then the minimal polynomial of u over $\mathbb{Q}$ is in fact a polynomial in $\mathbb{Z}[x]$.

§ 51
Kronecker's Theorem

In this paragraph, we prove an important theorem due to L. Kronecker which states that any polynomial over a field has a root in some extension field. It can be regarded as the fundamental theorem of field extensions. As might be expected from Kronecker's philosophical outlook, the proof is constructive: we do not merely prove the existence of such an extension in an unseen world; we actually describe what its elements are and how to add, multiply and invert them.

In our discussions concerning the roots of polynomials, we assumed, up to this point, that we are given: (1) a field K; (2) a polynomial $f(x)$ in $K[x]$; (3) an extension field E of K; (4) an element a of E which is a root of $f(x)$. But in many cases, we are given only a field K and a polynomial $f(x)$ in $K[x]$, and the problem is to find a root of $f(x)$. In more detail, the problem is to find a field E, an extension of K, and an element a in E such that $f(a) = 0$. Not only are we to find a, but we are also to find E, which is not given in advance. Kronecker's theorem tells us how to do this.

Let us consider a historical example, viz. the introduction of complex numbers into mathematics in the 18th and 19th centuries. Mathematicians had the field $\mathbb{R}$ of real numbers, and the polynomial $x^2 + 1 \in \mathbb{R}[x]$. This polynomial has no root in $\mathbb{R}$, because there is no real number whose square is -1. However, there were strong indications (for instance Cardan's formula for the roots of a cubic polynomial) that a root of this polynomial would be very welcome. What did mathematicians do, then? They invented a symbol $\sqrt{-1}$, which they perfectly knew not to be a real number, and considered the expressions $a + b\sqrt{-1}$, where $a,b \in \mathbb{R}$. These expressions were coined "complex numbers" (not a fortunate name, by the way). Two complex numbers $a + b\sqrt{-1}$ and $a' + b'\sqrt{-1}$ are regarded as equal if and only if $a = a'$ and $b = b'$. The sum of two complex numbers is defined in the obvious way. The product of two complex numbers $a + b\sqrt{-1}, c + d\sqrt{-1}$ is found from the naïve calculation

$$(a + b\sqrt{-1})(c + d\sqrt{-1}) = ac + ad\sqrt{-1} + b\sqrt{-1}c + b\sqrt{-1}d\sqrt{-1}$$

$$= ac + bd(\sqrt{-1})^2 + (ad + bc)\sqrt{-1}$$
$$= (ac - bd) + (ad + bc)\sqrt{-1},$$

where we interpret $(\sqrt{-1})^2$ as the real number -1. Thus $\sqrt{-1}$ is a computational device: we multiply complex numbers using the usual field properties of $\mathbb{R}$, and putting -1 for $(\sqrt{-1})^2$ wherever $(\sqrt{-1})^2$ occurs. The rigorous foundation for complex numbers as ordered pairs of real numbers, due to W. R. Hamilton, came in the middle of the 19th century, but there was nothing basically wrong in the "definition" of complex num-bers used by the earlier mathematicians. The field $\mathbb{C}$ were constructed in this way as the extension field $\mathbb{R}(i)$ of $\mathbb{R}$ having a root of the polynomial $x^2 + 1 \in \mathbb{R}[x]$. More specifically, the complex number $0 + 1\sqrt{-1}$ is a root of $x^2 + 1$.

Another example. Given the field $\mathbb{Q}$ and the polynomial $x^4 - 10x^2 + 1$ in $\mathbb{Q}[x]$, we wish to find a root of this polynomial. What can we do? As mentioned in Example 50.9(d), we invent a symbol a, subject it to the condition $a^4 - 10a^2 + 1 = 0$ and consider all expressions $c_0 + c_1a + c_2a^2 + c_3a^3$, as c_1,c_2,c_3,c_4 run independently over $\mathbb{Q}$. These expressions are new "numbers". These new "numbers" are multiplied using the usual field properties of $\mathbb{Q}$, and putting 0 for $a^4 - 10a^2 + 1$ wherever $a^4 - 10a^2 + 1$ occurs (equivalently, putting $10a^2 - 1$ for a^4 wherever a^4 occurs). The field $\mathbb{Q}(a)$ is constructed from $\mathbb{Q}$ and a as an extension field of $\mathbb{Q}$ having a root of the polynomial $x^4 - 10x^2 + 1 \in \mathbb{Q}[x]$. More specifically, the "number" $0 + 1a + 0a^2 + 0a^3$ is a root of $x^4 - 10x^2 + 1$.

It is now clear what to do in the general case. Given a field K and an irreducible polynomial $f(x)$ in $K[x]$, to find a root of $f(x)$, we invent a symbol u, subject it to the condition $f(u) = 0$ and consider the K-vector space with the K-basis $1,u,u^2, \ldots,u^{n-1}$, where $n = deg\ f(x)$ and $1,u,u^2, \ldots,u^{n-1}$ are computational symbols. We multiply the elements of this K-vector space by treating u as an indeterminate over K, and writing 0 for $f(u)$ wherever $f(u)$ occurs. The rigorous method of doing this is to consider the factor ring $K[x]/(f)$, as suggested by Theorem 50.6.

51.1 Theorem (Kronecker's theorem): *Let K be a field and $f(x)$ an irreducible polynomial in $K[x]$. Then there is an extension field E of K such that $f(x)$ has a root in E.*

Proof: Let $E = K[x]/(f)$, the factor ring of $K[x]$ modulo the principal ideal generated by $f(x)$ in $K[x]$. Since $K[x]$ is a principal ideal domain and $f(x)$ is irreducible in $K[x]$, the factor ring $E = K[x]/(f)$ is a field (Theorem 32.25).

The mapping
$$\varphi: K \longrightarrow E$$
$$k \to k + (f)$$
is a ring homomorphism because
$$(k_1 + k_2)\varphi = (k_1 + k_2) + (f) = (k_1 + (f)) + (k_2 + (f)) = k_1\varphi + k_2\varphi$$
and
$$(k_1 k_2)\varphi = k_1 k_2 + (f) = (k_1 + (f))(k_2 + (f)) = k_1\varphi \cdot k_2\varphi$$
for any $k_1, k_2 \in K$. Since $f(x)$ is irreducible in $K[x]$, it is not a unit in $K[x]$, thus $1\varphi = 1 + (f) \neq 0 + (f)$ and φ is one-to-one by Lemma 48.8. So φ is a field homomorphism. We identify K with its image $K\varphi$ in E. So we will write k instead of $k + (f)$ when $k \in K$. In this way, we regard K as a subfield of E and E as an extension field of K.

Let us write $u = x + (f) \in E$ for brevity. We claim that u is a root of $f(x)$. Indeed, if $f(x) = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n \in K[x]$, $b_n \neq 0$, then
$$f(u) = b_0 + b_1 u + b_2 u^2 + \cdots + b_n u^n$$
$$= (b_0 + (f)) + (b_1 + (f))(x + (f)) + (b_2 + (f))(x + (f))^2 + \cdots + (b_n + (f))(x + (f))^n$$
$$= (b_0 + (f)) + (b_1 + (f))(x + (f)) + (b_2 + (f))(x^2 + (f)) + \cdots + (b_n + (f))(x^n + (f))$$
$$= b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n + (f)$$
$$= f + (f)$$
$$= 0 + (f)$$
$$= 0 \in E$$
and so $u \in E$ is a root of $f(x)$. Thus E is an extension field of K containing a root of $f(x)$. (The identitification of K with $K\varphi \subseteq E$ amounts to writing k for $k + 0u + 0u^2 + \cdots + 0u^n \in E$ when $k \in K$.) □

Let us keep the notation of the preceding proof. Clearly $K(u) \subseteq E$. Also, any element of E has the form $c_0 + c_1 x + c_2 x^2 + \cdots + c_m x^m + (f)$, and thus equals
$$c_0 + c_1(x + (f)) + c_2(x + (f))^2 + \cdots + c_m(x + (f))^m$$
$$= c_0 + c_1 u + c_2 u^2 + \cdots + c_m u^m$$
and belongs to $K(u)$. So $E \subseteq K(u)$. This shows that $E = K(u)$ is a simple extension of K.

Now let $F = K(t)$ be another simple extension of K, generated by a root t in F of $f(x) \in K[x]$. By Theorem 50.6, we have the field isomorphisms

$$\begin{array}{ll} \alpha: K[x]/(f) \to K(u) & \beta: K[x]/(f) \to K(t) \\ g(x) + (f) \to g(u) & g(x) + (f) \to g(t) \end{array}$$

induced from the substitution homomorphisms

$$\begin{array}{ll} T_u: K[x] \to K(u) & T_t: K[x] \to K(t) \\ g(x) \to g(u) & g(x) \to g(t) \end{array}$$

(see Theorem 30.17). Hence

$$\begin{array}{l} \alpha^{-1}\beta: K(u) \to K(t) \\ g(u) \to g(t) \end{array}$$

is a field isomorphism: $K(u) \cong K(t)$. Besides, since $k\alpha = (k + (f))\alpha = kT_u = k$ and likewise $k\beta = k$ for all $k \in K$, the restriction of $\alpha^{-1}\beta$ to $K \subseteq K(u)$ is the identity mapping on K. We proved the following strengthening of Kronecker's theorem.

51.2 Theorem: *Let K be a field and let $f(x) \in K[x]$ be an irreducible polynomial in $K[x]$. Then there is a simple extension $K(u)$ of K such that $u \in K(u)$ is a root of $f(x)$. Moreover, if $K(t)$ is also a simple extension of K such that $t \in K(t)$ is a root of $f(x)$, then $K(u) \cong K(t)$ and in fact there is an isomorphism $\sigma: K(u) \to K(t)$ whose restriction to K is the identity mapping on K.* □

51.3 Definition: Let K be a field and let $f(x) \in K[x]$ be an irreducible polynomial in $K[x]$. Then a simple extension $K(u)$ of K, where u is a root of $f(x)$ (which field exists and is unique to within an isomorphism whose restriction to K is the identity mapping on K by Theorem 51.2), is called the field obtained by *adjoining a root of $f(x)$ to K*.

51.4 Remark: Let K be a field and let $f(x) \in K[x]$ be an irreducible polynomial in $K[x]$. Suppose $K(u)$ is the field obtained by adjoining a root u of $f(x)$ to K. Let c be the leading coefficient of $f(x)$. From Theorem 50.3,

we learn that $\frac{1}{c}f(x)$ is the minimal polynomial of u over K. Then it follows from Theorem 50.7 that $|K(u):K| = deg\ \frac{1}{c}f(x) = deg\ f(x)$: the degree over K of the field obtained by adjoining to K a root an irreducible polynomial $f(x) \in K[x]$ is equal to the degree of $f(x)$.

51.5 Theorem (Kronecker): *Let K be a field and let $f(x)$ be a polynomial in $K[x]\backslash K$ (not necessarily irreducible over K) with deg $f(x) = n$: Then there is an extension field E of K such that $f(x)$ has a root in E and $|E:K| \leqslant n$.*

Proof: From $f(x) \notin K$, we know that $f(x)$ is neither the zero polynomial nor a unit in $K[x]$. As $K[x]$ is a unique factorization domain, we can decompose $f(x)$ into irreducible polynomials, and adjoin a root of one of the irreducible divisors of $f(x)$ to K. The field E obtained in this way will have a root of (that irreducible divisor of $f(x)$, hence also of) $f(x)$. Moreover, $|E:K|$ will be equal to the degree of that irreducible divisor of $f(x)$, hence will be smaller than or equal to $deg\ f(x) = n$. □

51.6 Examples: (a) Consider the polynomial $f(x) = x^2 - 2 \in \mathbb{F}_5[x]$. It is irreducible over $\mathbb{F}_5$, for otherwise $f(x)$ would have a root in $\mathbb{F}_5$, whereas there is no element in $\mathbb{F}_5$ whose square is $2 \in \mathbb{F}_5$ (in the language of elementary number theory, $2 \in \mathbb{Z}$ is a quadratic nonresidue mod 5). Let us adjoin a root u of $f(x)$ to $\mathbb{F}_5$. The resulting field $\mathbb{F}_5(u)$ is an $\mathbb{F}_5$-vector space with an $\mathbb{F}_5$-basis $\{1,u\}$, and $u^2 = 2 \in \mathbb{F}_5$. Here are some sample computations in $\mathbb{F}_5(u)$:

$(4 + 2u)(3 + u) = 12 + 4u + 6u + 3u^2 = 12 + 10u + 3\cdot 2 = 2 + 0u + 1 = 3,$
$(3 + 2u)(2 + 4u) = 6 + 12u + 4u + 8u^2 = 1 + 2u + 4u + 3\cdot 2 = 7 + 6u = 2 + u.$

In view of the equation $u^2 = 2 \in \mathbb{F}_5$, we agree to write $\sqrt{2}$ in place of u in $\mathbb{F}_5(u)$. We keep in mind of course that $\sqrt{2}$ is just another name for our computational device u: here $\sqrt{2}$ is *not* the real number 1.414... whose square is the real number 2.

Let us express $(1 + 2\sqrt{2})(3 + \sqrt{2})$ and $(4 + \sqrt{2})^{-1}$ in terms of the $\mathbb{F}_5$-basis $\{1,\sqrt{2}\}$.

$(1 + 2\sqrt{2})(3 + \sqrt{2}) = 3 + \sqrt{2} + 6\sqrt{2} + 2\cdot 2 = (3 + 4) + (1 + 6)\sqrt{2} = 2 + 2\sqrt{2}$,

$$\frac{1}{4+\sqrt{2}} = \frac{1}{4+\sqrt{2}}\,\frac{4-\sqrt{2}}{4-\sqrt{2}} = \frac{4-\sqrt{2}}{16-2} = \frac{4-\sqrt{2}}{14} = \frac{4-\sqrt{2}}{4} = \frac{1}{4}(4 - \sqrt{2})$$

$$= 4(4-\sqrt{2}) = 16 - 4\sqrt{2} = 1 + \sqrt{2}.$$

Check: $(4 + \sqrt{2})(1 + \sqrt{2}) = 4 + 4\sqrt{2} + \sqrt{2} + 2 = 6 + 5\sqrt{2} = 1$.

Note that $\varphi: \mathbb{F}_5(\sqrt{2}) \to \mathbb{F}_5(\sqrt{2})$ is an automorphism of $\mathbb{F}_5(\sqrt{2})$, because
$a + b\sqrt{2} \to a - b\sqrt{2}$

$$\begin{aligned}[(a + b\sqrt{2}) + (c + d\sqrt{2})]\varphi &= [(a + c) + (b + d)\sqrt{2}]\varphi \\ &= (a + c) - (b + d)\sqrt{2} \\ &= (a - b\sqrt{2}) + (c - d\sqrt{2}) \\ &= (a + b\sqrt{2})\varphi + (c + d\sqrt{2})\varphi\end{aligned}$$

and
$$\begin{aligned}[(a + b\sqrt{2})(c + d\sqrt{2})]\varphi &= [(ac + 2bd) + (ad + bc)\sqrt{2}]\varphi \\ &= (ac + 2bd) - (ad + bc)\sqrt{2} \\ &= (ac + 2(-b)(-d)) + (a(-d) + (-b)c)\sqrt{2} \\ &= (a - b\sqrt{2})(c - d\sqrt{2}) \\ &= (a + b\sqrt{2})\varphi\cdot(c + d\sqrt{2})\varphi\end{aligned}$$

for all $a + b\sqrt{2}, c + d\sqrt{2} \in \mathbb{F}_5(\sqrt{2})$; and φ is clearly onto and $Ker\ \varphi \neq \mathbb{F}_5(\sqrt{2})$.
By the binomial theorem (Theorem 29.16),

$$\begin{aligned}(a + b\sqrt{2})^5 &= a^5 + 5a^4b\sqrt{2} + 10a^3b^2 2 + 10a^2b^3 2\sqrt{2} + 5ab^4 4 + b^5 4\sqrt{2} \\ &= a^5 + 4b^5\sqrt{2} = a + 4b\sqrt{2} = a - b\sqrt{2}\end{aligned}$$

for all $a + b\sqrt{2} \in \mathbb{F}_5(\sqrt{2})$. Thus φ can also be described as

$$\varphi: \mathbb{F}_5(\sqrt{2}) \to \mathbb{F}_5(\sqrt{2}).$$
$$t \to t^5$$

(b) The polynomial $g(x) = x^2 - 3 \in \mathbb{F}_5[x]$, too, is irreducible over $\mathbb{F}_5$ ($3 \in \mathbb{Z}$ is a quadratic nonresidue mod 5). Adjoining a root $\sqrt{3}$ of $g(x)$ to $\mathbb{F}_5$, we obtain the field $\mathbb{F}_5(\sqrt{3})$, which is an $\mathbb{F}_5$-vector space with a basis $\{1, \sqrt{3}\}$ over $\mathbb{F}_5$, and $(\sqrt{3})^2 = 3 \in \mathbb{F}_5$. We do not forget, of course, that $\sqrt{3}$ is a computational symbol only, and *not* the real number 1.732... whose square is $3 \in \mathbb{R}$. In $\mathbb{F}_5(\sqrt{3})$, we have

$(3 + 2\sqrt{3})(1 + 4\sqrt{3}) = 3 + 12\sqrt{3} + 2\sqrt{3} + 8\cdot 3 = 27 + 14\sqrt{3} = 2 + 4\sqrt{3}$,
$(2 + 3\sqrt{3})(2 + 4\sqrt{3}) = 4 + 8\sqrt{3} + 6\sqrt{3} + 12\cdot 3 = 4 + 3\sqrt{3} + \sqrt{3} + 36 = 4\sqrt{3}$,

$$\frac{1}{1+3\sqrt{3}} = \frac{1}{1+3\sqrt{3}}\,\frac{1-3\sqrt{3}}{1-3\sqrt{3}} = \frac{1-3\sqrt{3}}{1-27} = \frac{1+2\sqrt{3}}{4} = \frac{1}{4}(1 + 2\sqrt{3})$$

$$= 4(1 + 2\sqrt{3}) = 4 + 3\sqrt{3}.$$

As $8 = 3$ in $\mathbb{F}_5$, we may also write $\sqrt{8}$ for $\sqrt{3}$, with the understanding that $\sqrt{8} \in \mathbb{F}_5(\sqrt{3})$ is a computational device satisfying $(\sqrt{8})^2 = 8 = 3$. Here $\sqrt{8}$ is *not* the real number 2.828... whose square is $8 \in \mathbb{R}$. We might be tempted to write $\sqrt{8} = \sqrt{4 \cdot 2} = 2\sqrt{2}$. For the time being, this is not legitimate: as $\sqrt{8} \in \mathbb{F}_5(\sqrt{3})$ and $2\sqrt{2} \in \mathbb{F}_5(\sqrt{2})$ are in different fields, and not in their intersection $\mathbb{F}_5$, it is not meaningful to write $\sqrt{8} = 2\sqrt{2}$.

However, this suggests that $\psi: \mathbb{F}_5(\sqrt{3}) \to \mathbb{F}_5(\sqrt{2})$ might be an interesting

$$a + b\sqrt{3} \to a + 2b\sqrt{2}$$

mapping. Indeed,

$$\begin{aligned}
[(a + b\sqrt{3}) + (c + d\sqrt{3})]\psi &= [(a + c) + (b + d)\sqrt{3}]\psi \\
&= (a + c) + 2(b + d)\sqrt{2} \\
&= (a + 2b\sqrt{2}) + (c + 2d\sqrt{2}) \\
&= (a + b\sqrt{3})\psi + (c + d\sqrt{3})\psi
\end{aligned}$$

and

$$\begin{aligned}
[(a + b\sqrt{3})(c + d\sqrt{3})]\psi &= [(ac + 3bd) + (ad + bc)\sqrt{3}]\psi \\
&= (ac + 3bd) + 2(ad + bc)\sqrt{2} \\
&= (ac + 2 \cdot 2b \cdot 2d) + (a \cdot 2d + 2b \cdot c)\sqrt{2} \\
&= (a + 2b\sqrt{2})(c + 2d\sqrt{2}) \\
&= (a + b\sqrt{3})\psi \cdot (c + d\sqrt{3})\psi
\end{aligned}$$

for all $a + b\sqrt{3}, c + d\sqrt{3} \in \mathbb{F}_5(\sqrt{3})$, thus ψ is a ring homomorphism. As it is clearly one-to-one and onto, ψ is a field isomorphism. Hence $\mathbb{F}_5(\sqrt{3})$ and $\mathbb{F}_5(\sqrt{2})$ are isomorphic fields. We *identify* these two fields by the isomorphism ψ, i.e., by declaring $a + b\sqrt{3} = a + 2b\sqrt{2}$ for all $a,b \in \mathbb{F}_5$. Then, but only then can we write $\sqrt{3} = 2\sqrt{2}$.

We could identify these fields by declaring $a + b\sqrt{3} = a - 2b\sqrt{2}$ for all a,b in $\mathbb{F}_5$, which amounts to identifying them by the isomorphism $\varphi\psi: \mathbb{F}_5(\sqrt{3}) \to \mathbb{F}_5(\sqrt{2})$. How we identify them is not important, but we must consistently use one and the same identification.

When we identify $\mathbb{F}_5(\sqrt{3})$ and $\mathbb{F}_5(\sqrt{2})$ by declaring $a + b\sqrt{3} = a + 2b\sqrt{2}$ for all $a,b \in \mathbb{F}_5$, we can no longer interpret $\sqrt{18}$, for example, merely as a computational device whose square is $18 \in \mathbb{F}_5$, for there are *two* elements in $\mathbb{F}_5(\sqrt{3}) = \mathbb{F}_5(\sqrt{2})$ whose squares are 18, viz. $2\sqrt{2}$ and $-2\sqrt{2} = 3\sqrt{2}$. We must specify which of $2\sqrt{2}, 3\sqrt{2}$ we mean by $\sqrt{18}$. Otherwise we might commit such mistakes as

$$3\sqrt{2} = \sqrt{9 \cdot 2} = \sqrt{18} = \sqrt{9 \cdot 2} = \sqrt{4 \cdot 2} = 2\sqrt{2} \qquad \text{in } \mathbb{F}_5(\sqrt{2})$$

which resembles the mistake

$$-7 = \sqrt{(-7)^2} = \sqrt{49} = 7 \qquad \text{in } \mathbb{R}.$$

In $\mathbb{R}$, there are two numbers whose squares are 49, namely 7 and -7, and $\sqrt{49}$ is understood to be the positive of the numbers 7, -7. Thus when we write $\sqrt{49}$, we specify which of 7, -7 we mean by $\sqrt{49}$. This prevents the mistake $-7 = \sqrt{(-7)^2}$. In $\mathbb{F}_5(\sqrt{2})$, specifying $2\sqrt{2}$ or $3\sqrt{2}$ as $\sqrt{18}$ prevents the mistake $3\sqrt{2} = 2\sqrt{2}$.

Exercises

1. Adjoin a root u of $x^3 + 2x^2 - 2 \in \mathbb{Q}[x]$ to $\mathbb{Q}$ and construct the field $\mathbb{Q}(u)$. Express $(u^2 + u - 1)(u^2 + 2u - 5)$, $(u^2 - 3u + 1)/(u^2 + 2u + 3)$, $(u^4 + u^3)(u^3 - 1)$ in terms of the $\mathbb{Q}$-basis $1, u, u^2$ of $\mathbb{Q}(u)$.

2. Find all monic irreducible polynomials in $\mathbb{F}_5[x]$ of degree two (aside from $x^2 - 2$ and $x^2 - 3$, there are eight of them). Adjoining a root u of these polynomials to $\mathbb{F}_5$, construct eight fields $\mathbb{F}_5(u)$ of 25 elements. Prove that each of these fields is isomorphic to $\mathbb{F}_5(\sqrt{2})$.

3. Prove that $\mathbb{F}_5^\times$ and $\mathbb{F}_5(\sqrt{2})^\times$ are cyclic.

4. Find a field K of nine elements and show that $K^\times$ is cyclic.

5. Prove the following statements.

   (a) $f(x) = x^2 + x + 1 \in \mathbb{F}_2[x]$ is irreducible over $\mathbb{F}_2$.

   (b) $g(x) = x^4 + x + 1 \in \mathbb{F}_2[x]$ is irreducible over $\mathbb{F}_2$.

Let i be a root of $f(x)$ and u a root of $g(x)$.

   (c) $h(x) = x^2 + ix + 1 \in \mathbb{F}_2(i)[x]$ is irreducible over $\mathbb{F}_2(i)$.

Let t be a root of $h(x) \in \mathbb{F}_2(i)[x]$.

   (d) $\mathbb{F}_2(i)(t)^\times$ and $\mathbb{F}_2(u)^\times$ are cyclic.

   (e) $\mathbb{F}_2(i)(t) \cong \mathbb{F}_2(u)$.

§52
Finite Fields

We have seen some examples of finite fields, i.e., fields with finitely many elements. In this paragraph, we want to discuss some properties of finite fields.

In modern times, it is customary to treat finite fields after the presentation of Galois theory. Our approach to finite fields will be more elementary and more concrete than usual. We hope this will prepare the way to a better understanding of Galois theory. See also Example 54.18(c) and Theorem 54.26.

We begin by restricting the order of a finite field to prime powers.

52.1 Lemma: *Let q be a natural number and K a field with q elements. Then $q = p^n$ for some prime number p and for some natural number n.*

Proof: Let K be a field with q elements. The prime subfield of K cannot be (isomorphic to) $\mathbb{Q}$, for then K would contain infinitely many elements. Hence the prime subfield of K is (isomorphic to) $\mathbb{F}_p$ for some prime number p. We consider K as an $\mathbb{F}_p$-vector space. The dimension of K over $\mathbb{F}_p$ must be finite, say $|K:\mathbb{F}_p| = n \in \mathbb{N}$. Let $\{k_1, k_2, \ldots, k_n\}$ be an $\mathbb{F}_p$-basis of K. Then K consists of the elements

$$a_1k_1 + a_2k_2 + \cdots + a_nk_n$$

as $a_1, a_2, \ldots, a_n$ run independently through $\mathbb{F}_p$, and

$$a_1k_1 + a_2k_2 + \cdots + a_nk_n \neq b_1k_1 + b_2k_2 + \cdots + b_nk_n$$

whenever $(a_1, a_2, \ldots, a_n) \neq (b_1, b_2, \ldots, b_n)$. Hence there are p possible choices for each of $a_1, a_2, \ldots, a_n$ and there are precisely $pp\ldots p = p^n$ elements in K.

Thus the condition $q = p^n$ is a necessary condition for the existence of a field with q elements. One of our main goals in this paragraph is to show that it is also a sufficient condition.

By the proof of Lemma 52.1, we know that a field with p^n elements is of characteristic p. We prove two lemmas about (not necessarily finite) fields of prime characteristic.

52.2 Lemma: *Let K be a field of characteristic $p \neq 0$. Then*

$$(a + b)^p = a^p + b^p \qquad \text{and} \qquad (ab)^p = a^p b^p$$

for all $a,b \in K$.

Proof: We use the binomial theorem (Theorem 29.16). Here p is a prime number and the binomial coefficients $\binom{p}{k}$ are divisible by p when $k = 1,2,\ldots,p-1$: note that $p!$ is divisible by p, so $k!(p-k)!\binom{p}{k}$ is divisible by p, but $k!(p-k)!$ is relatively prime to p, so p divides $\binom{p}{k}$ by Theorem 5.12. Then, for any $a,b \in K$, we have

$$(a + b)^p = a^p + \sum_{k=1}^{p-1} \binom{p}{k} a^{p-k} b^k + b^p = a^p + \sum_{k=1}^{p-1} 0 + b^p = a^p + b^p$$

since $p \mid \binom{p}{k}$ and *char* $K = p$ imply that $\binom{p}{k} a^{p-k} b^k = 0$ for $k = 1,2,\ldots,p-1$. This proves $(a + b)^p = a^p + b^p$. The claim $(ab)^p = a^p b^p$ follows from Lemma 8.14(1). □

Lemma 52.2 states that the mapping $\sigma : K \to K$, $a \to a^p$ is a field homomorphism (clearly $1^p = 1 \neq 0$). By induction on m, we obtain

$$(a_1 + a_2 + \cdots + a_m)^p = a_1^p + a_2^p + \cdots + a_m^p$$

for any m elements $a_1, a_2, \ldots, a_m$ of a field of prime characteristic p.

52.3 Lemma: *Let K be a field of characteristic $p \neq 0$ and $n \in \mathbb{N}$. Then,*

$$(a + b)^{p^n} = a^{p^n} + b^{p^n}$$

for any $a,b \in K$.

Proof: We make induction on n. The claim is established for $n = 1$ in Lemma 52.2. If the assertion is true for $n = k$, then, for any $a,b \in K$,

$$(a + b)^{p^{k+1}} = [(a + b)^{p^k}]^p = [a^{p^k} + b^{p^k}]^p = (a^{p^k})^p + (b^{p^k})^p = a^{p^{k+1}} + b^{p^{k+1}}$$

and it is true for $n = k + 1$ also. Hence it is true for all $n \in \mathbb{N}$. □

52.4 Lemma: *Let $q \in \mathbb{N}$ and let K be a field with q elements.*
(1) $a^{q-1} = 1$ *for all* $a \in K^\times$.
(2) $a^q = a$ *for all* $a \in K$.

(3) $x^q - x = \prod_{a \in K} (x - a)$ *in* $K[x]$.

(4) *Let $f(x)$ be a nonzero polynomial of degree d in $K[x]$. If $f(x) | (x^q - x)$ in $K[x]$, then $f(x)$ has exactly d roots in K, and these roots are pairwise distinct.*

Proof: (1) $K^\times$ is a multiplicative group of order $|K^\times| = |K \setminus \{0\}| = q - 1$. Hence $a^{q-1} = 1$ for any $a \in K^\times$.

(2) This follows from (1) if $a \neq 0$ and from $0^q = 0$ if $a = 0$.
(3) Any element a of K is a root of the polynomial $x^q - x \in K[x]$ by part (2). Thus both $x^q - x$ and $\prod_{a \in K} (x - a)$ are monic polynomials, in $K[x]$, of degree q having all the q elements of K as roots. If $x^q - x$ were not equal to $\prod_{a \in K} (x - a)$, then $(x^q - x) - \prod_{a \in K} (x - a)$ would be a nonzero polynomial of degree less than q having at least q distinct roots, contrary to Theorem 35.7. So $x^q - x = \prod_{a \in K} (x - a)$

(4) We put $x^q - x = f(x)g(x)$, with $g(x) \in K[x]$. Then *deg* $g(x) = q - d$. The roots of $x^q - x$ are pairwise distinct by part (3) and, since any root of $f(x)$ is also a root of $x^q - x$, we see that the roots of $f(x)$, too, are pairwise distinct. Likewise the roots of $g(x)$ are pairwise distinct. Now $g(x)$ has at most $q - d$ roots in K (Theorem 35.7). If $f(x)$ had r roots in K and $r < d$, then $x^q - x = f(x)g(x)$ would have at most $r + (q - d) < q$ roots in K, contrary to the fact that all q elements of K are roots of $x^q - x$. Thus $f(x)$ has at least d roots in K. But it can have at most d roots in K by Theorem 35.7. Hence $f(x)$ has exactly d roots in K. □

52.5 Lemma: *Let L/K be a field extension and assume that K has q elements, $q \in \mathbb{N}$. Let b be an element of L. Then $b \in K$ if and only if $b^q = b$.*

Proof: $b \in K$ if and only if b is a root of $\prod_{a \in K} (x - a)$, so if and only if b is a root of $x^q - x$, so if and only if $b^q = b$. □

The last two lemmas will now be employed to get information about the subfields of a finite field. If $K_1 \subseteq K_2$ are finite fields, with p^{m_1} and p^{m_2} elements, respectively, then $K_1^\times$ is a subgroup of $K_2^\times$, hence $p^{m_1} - 1 = |K_1^\times|$ divides $|K_2^\times| = p^{m_2} - 1$ by Lagrange's theorem. We proceed to show that this happens if and only if m_1 divides m_2.

52.6 Lemma: *Let $m,n \in \mathbb{N}$ and put $d = (m,n)$.*
(1) For any $k \in \mathbb{N}$, we have $(k^m - 1, k^n - 1) = k^d - 1$.
(2) If K is any field and x an indeterminate over K, then, in the unique factorization domain $K[x]$, we have $(x^m - 1, x^n - 1) \approx x^d - 1$.

Proof: (1) We put $e = (k^m - 1, k^n - 1)$. Since

$$k^m - 1 = (k^d - 1)((k^d)^{(m/d)-1} + (k^d)^{(m/d)-2} + \cdots + k^d + 1),$$

we have $k^d - 1 | k^m - 1$. Likewise $k^d - 1 | k^n - 1$ and so $k^d - 1 | e$. On the other hand, $k^m \equiv 1 \pmod{e}$, so $\overline{k}^m = \overline{1}$ in $\mathbb{Z}_e^\times$, so $o(\overline{k}) | m$. Likewise $o(\overline{k}) | n$, so $o(\overline{k}) | d$, so $\overline{k}^d = \overline{1}$ in $\mathbb{Z}_e^\times$, so $k^d \equiv 1 \pmod{e}$, so $e | k^d - 1$. From $k^d - 1 | e$ and $e | k^d - 1$, we obtain $e = k^d - 1$, as claimed.

(2) We put $f(x) = (x^m - 1, x^n - 1)$. Since

$$x^m - 1 = (x^d - 1)((x^d)^{(m/d)-1} + (x^d)^{(m/d)-2} + \cdots + x^d + 1),$$

we have $x^d - 1 | x^m - 1$ in $K[x]$. Likewise $x^d - 1 | x^n - 1$ and so $x^d - 1 | f(x)$. On the other hand, $f(x) | x^m - 1$, so $(x + (f))^m = x^m + (f) = 1 + (f)$ in $K[x]/(f(x))$, hence $x + (f)$ is a unit in $K[x]/(f)$ and the order of $x + (f) \in (K[x]/(f))^\times$ is divisible by m, likewise by n, and therefore by d. Thus $x^d + (f) = (x + (f))^d = 1 + (f)$, and $f(x) | x^d - 1$ in $K[x]$. From $x^d - 1 | f(x)$ and $f(x) | x^d - 1$, we get $f(x) \approx x^d - 1$, as claimed. □

52.7 Lemma: *Let $m,n,p \in \mathbb{N}$ and let K be a field and x an indeterminate over K.*

(1) *For any $k \in \mathbb{N}$, we have $k^m - 1 | k^n - 1$ if and only if $m|n$.*

(2) *In the polynomial ring $K[x]$, we have $x^m - 1 | x^n - 1$ if and only if $m|n$.*

(3) *In the polynomial ring $K[x]$, we have $x^{p^m} - x \,|\, x^{p^n} - x$ if and only if $m|n$.*

Proof: (1) $k^m - 1 | k^n - 1$ if and only if $(k^m - 1, k^n - 1) = k^m - 1$, so if and only if $k^{(m,n)} - 1 = k^m - 1$, so if and only if $(m,n) = m$, so if and only if $m|n$.

(2) $x^m - 1 | x^n - 1$ in $K[x]$ if and only if $(x^m - 1, x^n - 1) \approx x^m - 1$, so if and only if $x^{(m,n)} - 1 \approx x^m - 1$, so if and only if $x^{(m,n)} - 1 = x^m - 1$, so if and only if $(m,n) = m$, so if and only if $m|n$.

(3) We have $x^{p^m} - x \,|\, x^{p^n} - x$ if and only if $x^{p^m-1} - 1 \,|\, x^{p^n-1} - 1$, so if and only if $p^m - 1 \,|\, p^n - 1$ by part (2), so if and only if $m|n$ by part (1). □

52.8 Theorem: *Let K be a field with p^n elements (p prime). Then K has a subfield with p^m elements if and only if $m|n$. In this case, there is exactly one subfield of K with p^m elements. This subfield is*

$$\{a \in K : a^{p^m} = a\}.$$

Proof: As noted earlier, if K has a subfield H with p^m elements, then $H^\times$ is a subgroup of $K^\times$, so $p^m - 1 = |H^\times|$ divides $|K^\times| = p^n - 1$ by Lagrange's theorem. From $p^m - 1 | p^n - 1$, we get $m|n$ by Lemma 52.7(1).

Suppose now $m|n$. We want to show that K has a subfield with p^m elements. Lemma 52.5 leads us to consider the set of all elements a in K satisfying $a^{p^m} = a$. So we put $K_1 = \{a \in K : a^{p^m} = a\}$. Then K_1 is not empty and, for any $a,b \in K_1$, we have

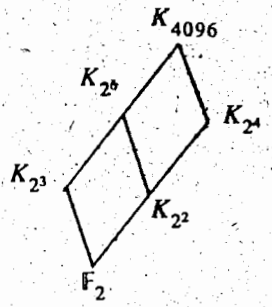
$$(a+b)^{p^m} = a^{p^m} + a^{p^m} = a + b, \qquad \text{so } a + b \in K_1 \qquad \text{(Lemma 52.3)},$$
$$(-b)^{p^m} = (-1b)^{p^m} = (-1)^{p^m} b^{p^m} = (-1)b, \qquad \text{so } -b \in K_1 \qquad \text{(even when } p = 2),$$
$$(ab)^{p^m} = a^{p^m} b^{p^m} = ab, \qquad \text{so } ab \in K_1,$$
$$(1/b)^{p^m} = 1/b^{p^m} = 1/b, \qquad \text{so } 1/b \in K_1 \qquad \text{(if } b \neq 0).$$

Thus K_1 is a subfield of K. We now show that K_1 has exactly p^m elements. Since $m|n$, we have $x^{p^m} - x \,|\, x^{p^n} - x$ in $K[x]$ (Lemma 52.7(3)). Thus the polynomial $x^{p^m} - x$ has exactly p^m roots and these are pairwise distinct

(Lemma 52.4(4)) and the roots of $x^{p^m} - x$ are precisely the elements in K_1. Hence K_1 has indeed p^m elements.

This proves that K has a subfield K_1 with p^m elements whenever $m|n$. Moreover, there is only one subfield with p^m elements, for if K_2 is a subfield of K and $|K_2| = p^m$, then any element b of K_2 satisfies $b^{p^m} = b$ by Lemma 52.5, so $K_2 \subseteq K_1$, so $K_2 = K_1$. The proof is complete. □

As an illustration of Theorem 52.8, assume that K_{4096} is a field with $4096 = 2^{12}$ elements. Then all subfields of K_{4096} are as the figure below, where K_q denotes a field with q elements.



In particular, assuming the existence of a field with 4096 elements, we can conclude the existence of a field with 2^1, 2^2, 2^3, 2^4, 2^6 elements, too. However, we do not know whether a field with 4096 elements really exists, so the foregoing argument is very weak. It is in fact true that there is a field with p^n elements, for any prime number p and for any natural number n. We wish to prove this assertion. We need some results from elementary number theory.

*

* *

In the following, we use the notation $\sum_{d|n} a_d$. This means that $n \in \mathbb{N}$ and that we take a sum of terms a_d as d ranges through the positive divisors of n, including 1 and n. For instance $\sum_{d|12} a_d = a_1 + a_2 + a_3 + a_4 + a_6 + a_{12}$ and

$\sum_{d|15} = a_1 + a_3 + a_5 + a_{15}$. We cleary have $\sum_{d|n} a_d = \sum_{d|n} a_{n/d}$. The notations $\prod_{d|n} a_d$ and $\bigcup_{d|n} S_d$ will have similar meanings.

52.9 Lemma: *Let φ be Euler's function. Then, for any natural number n,*

$$\sum_{d|n} \varphi(d) = n.$$

Proof: For any $k \in \mathbb{N}$, $\varphi(k)$ is defined to be the number of positive integers less than (on equal to) k that are relatively prime to k. The greatest common divisor of any integer in $\{1,2, \ldots,n\}$ with n is a positive divisor d of n. Hence we have

$$\{1,2,\ldots,n\} = \bigcup_{d|n} S_d, \qquad \text{where } S_d = \{k \in \mathbb{N}: k \leqslant n \text{ and } (k,n) = d\}.$$

Counting the number of elements, we get $n = |\{1,2, \ldots,n\}| = \sum_{d|n} |S_d|$. Here

$$\begin{aligned} S_d &= \{k \in \mathbb{N}: k \leqslant n \text{ and } (k,n) = d\} \\ &= \{k \in \mathbb{N}: d|k, k \leqslant n \text{ and } (k,n) = d\}. \\ &= \{k \in \mathbb{N}: k = db \text{ for some } b \in \mathbb{N}, k \leqslant n \text{ and } (k,n) = d\} \\ &= \{db \in \mathbb{N}:, db \leqslant n \text{ and } (db,n) = d\} \\ &= \{db \in \mathbb{N}:, db \leqslant n \text{ and } (db,d\tfrac{n}{d}) = d\} \\ &= \{db \in \mathbb{N}: 1 \leqslant b \leqslant \tfrac{n}{d} \text{ and } (b,\tfrac{n}{d}) = 1\}. \end{aligned}$$

Thus $|S_d|$ is the number of positive integers b such that $1 \leqslant b \leqslant n/d$ and $(b, (n/d)) = 1$, and this number is $\varphi(n/d)$ by definition. We then obtain

$$n = \sum_{d|n} |S_d| = \sum_{d|n} \varphi(n/d) = \sum_{d|n} \varphi(d). \qquad \square$$

For ease in formulation of the next lemma, we introduce some terminology. Let $m \in \mathbb{N}$. A *complete residue system mod m* is defined to be a set of m integers such that one and only one of them is congruent to each one of $1,2, \ldots,m$. Thus a complete residue system mod m is a set

$\{r_1,r_2,\ldots,r_m\} \subseteq \mathbb{Z}$ such that the residue classes mod m of $r_1,r_2,\ldots,r_m$ make up $\mathbb{Z}_m$. In particular, r_i are then mutually incongruent mod m (and, *a fortiori*, mutually distinct). If $r_1,r_2,\ldots,r_m$ are integers mutually incongruent mod m, then $\{r_1,r_2,\ldots,r_m\}$ is a complete residue system mod m. Also, if any integer is congruent, modulo m, to one of the integers $r_1,r_2,\ldots,r_m$, then $\{r_1,r_2,\ldots,r_m\}$ is a complete residue system mod m.

A *reduced residue system mod m* is defined to be a set of $\varphi(m)$ integers such that one and only one of them is congruent to each one of the integers among $1,2,\ldots,m$ that are relatively prime to m. Thus a reduced residue system mod m is a set $\{a_1,a_2,\ldots,a_{\varphi(m)}\} \subseteq \mathbb{Z}$ such that the residue classes mod m of $a_1,a_2,\ldots,a_{\varphi(m)}$ make up $\mathbb{Z}_m^*$. In particular, a_i are then mutually incongruent mod m (and, *a fortiori*, mutually distinct). If $a_1,a_2,\ldots,a_{\varphi(m)}$ are integers relatively prime to m and mutually incongruent mod m, then $\{a_1,a_2,\ldots,a_{\varphi(m)}\}$ is a reduced residue system mod m. Also, if any integer that is relatively prime to m is congruent, modulo m, to one of the integers $a_1,a_2,\ldots,a_{\varphi(m)}$, then $\{a_1,a_2,\ldots,a_{\varphi(m)}\}$ is a reduced residue system mod m.

52.10 Lemma: *Let φ be Euler's function. Let $m,n \in \mathbb{N}$ and $(m,n) = 1$.*

(1) *If $\{r_1,r_2,\ldots,r_m\} \subseteq \mathbb{Z}$ is a complete residue system mod m and if $\{s_1,s_2,\ldots,s_n\} \subseteq \mathbb{Z}$ is a complete residue system mod n, then*

$$\{ms_i + nr_j : i = 1,2,\ldots,m, j = 1,2,\ldots,n\} \subseteq \mathbb{Z}$$

is a complete residue system mod mn.

(2) *If $\{a_1,a_2,\ldots,a_{\varphi(m)}\} \subseteq \mathbb{Z}$ is a reduced residue system mod m and if $\{b_1,b_2,\ldots,b_{\varphi(n)}\} \subseteq \mathbb{Z}$ is a reduced residue system mod n, then*

$$\{ma_i + nb_j : i = 1,2,\ldots,\varphi(m), j = 1,2,\ldots,\varphi(n)\} \subseteq \mathbb{Z}$$

is a reduced residue system mod mn.

(3) $\varphi(mn) = \varphi(m)\varphi(n)$.

Proof: (1) It will be sufficient to show that any two distinct of the mn numbers $ms_i + nr_j$ are incongruent modulo mn. Indeed, if

$$ms_i + nr_j \equiv ms_{i'} + nr_{j'} \pmod{mn},$$

then
$$ms_i + nr_j \equiv ms_{i'} + nr_{j'} \pmod{m} \text{ and } ms_i + nr_j \equiv ms_{i'} + nr_{j'} \pmod{n}$$
$$nr_j \equiv nr_{j'} \pmod{m} \text{ and } ms_i \equiv ms_{i'} \pmod{n}$$
$$r_j \equiv r_{j'} \pmod{m} \text{ and } s_i \equiv s_{i'} \pmod{n}$$
$$r_j = r_{j'} \text{ and } s_i = s_{i'}$$

$$ms_i + nr_j = ms_{i'} + nr_{j'}.$$

(2) Let us take a complete residue system $\{r_1,r_2,\ldots,r_m\}$ mod m such that $\{a_1,a_2,\ldots,a_{\varphi(m)}\} \subseteq \{r_1,r_2,\ldots,r_m\}$ and a complete residue system $\{s_1,s_2,\ldots,s_m\}$ mod n such that $\{b_1,b_2,\ldots,b_{\varphi(n)}\} \subseteq \{s_1,s_2,\ldots,s_n\}$. We have $\{a_1,a_2,\ldots,a_{\varphi(m)}\} = \{r_j : j = 1,2,\ldots,m, (r_j,m) = 1\}$ and $\{b_1,b_2,\ldots,b_{\varphi(n)}\} = \{s_i : i = 1,2,\ldots,n, (s_i,n) = 1\}$. Now $\{ms_i + nr_j : j = 1,2,\ldots,m, j = 1,2,\ldots,n\}$ is a complete residue system mod mn. So it will be sufficient to show that $ms_i + nr_j$ is relatively prime to mn if and only if $(s_i,n) = 1$ and $(r_j,m) = 1$.

If $(s_i,n) > 1$, then (s_i,n) divides both $ms_i + nr_j$, and mn, so (s_i,n) divides $(ms_i + nr_j, mn)$ and $(ms_i + nr_j, mn) > 1$. Likewise $(r_j,m) > 1$ implies that $(ms_i + nr_j, mn) > 1$.

On the other hand, if $(s_i,n) = 1$ and $(r_j,m) = 1$, then $(ms_i + nr_j,mn) = 1$. For otherwise $(ms_i + nr_j, mn)$ would be divisible by a prime number p. Then we would have $p|mn$, so $p|m$ or $p|n$. Without loss of generality, assume $p|m$. Also $p|ms_i + nr_j$, so $p|nr_j$. Since $p|m$ and $(m,n) = 1$, we would get $(p,n) = 1$. Then $p|nr_j$ and $(p,n) = 1$ would give $p|r_j$ and p would divide (r_j,m), contrary to $(r_j,m) = 1$. So $(s_i,n) = 1$ and $(r_j,m) = 1$ implies $(ms_i + nr_j, mn) = 1$.

(3) From part (2), we learn that a reduced residue system modulo mn has $\varphi(m)\varphi(n)$ elements. Hence $\varphi(mn) = \varphi(m)\varphi(n)$ whenever m and n are relatively prime. □

It follows by induction on k that $\varphi(m_1m_2\ldots m_k) = \varphi(m_1)\varphi(m_2)\ldots\varphi(m_k)$ for all natural numbers $m_1,m_2,\ldots,m_k$ that are pairwise relatively prime. In particular, if $n \in \mathbb{N}$ and $n = p_1^{a_1}p_2^{a_2}\ldots p_k^{a_k}$ is the canonical decomposition of n into prime numbers, then $\varphi(n) = \varphi(p_1^{a_1})\varphi(p_2^{a_2})\ldots\varphi(p_k^{a_k})$.

Now it is easy to find $\varphi(p^a)$ in closed form if p is prime: among the p^a integers $1,2,\ldots,p^a$, exactly p^{a-1} of them, namely

$$p1, p2, \ldots, pp^{a-1}$$

are not relatively prime to p, so exactly $p^a - p^{a-1}$ of them are relatively prime to p. This means $\varphi(p^a) = p^a - p^{a-1}$. We can also write $\varphi(p^a)$ $= p^a\left(1 - \frac{1}{p}\right)$.

Therefore, if $n \in \mathbb{N}$, $n > 1$ and $n = p_1^{a_1}p_2^{a_2}\dots p_k^{a_k}$ is the canonical decomposition of n into prime numbers, then

$$\begin{aligned}\varphi(n) &= (p_1^{a_1} - p_1^{a_1-1})(p_2^{a_2} - p_2^{a_2-1})\dots(p_k^{a_k} - p_k^{a_k-1}) \\ &= p_1^{a_1-1}p_2^{a_2-1}\dots p_k^{a_k-1}(p_1 - 1)(p_2 - 1)\dots(p_k - 1) \\ &= p_1^{a_1}p_2^{a_2}\dots p_k^{a_k}\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\dots\left(1 - \frac{1}{p_k}\right) \\ &= n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\dots\left(1 - \frac{1}{p_k}\right).\end{aligned}$$

Expanding the last expresion, we find

$$\varphi(n) = n - \left(\frac{n}{p_1} + \frac{n}{p_2} + \dots + \frac{n}{p_k}\right) + \left(\frac{n}{p_1p_2} + \frac{n}{p_1p_3} + \dots + \frac{n}{p_{k-1}p_k}\right) + \dots + (-1)^k\left(\frac{n}{p_1p_2\dots p_k}\right).$$

Thus $\varphi(n)$ is equal to a sum of terms of the form $\mp \frac{n}{d}$, where d is a product of distinct prime divisors of n, and the sign is + or − according as the number of prime divisors is even or odd. Thus we can write

$$\varphi(n) = \sum_{d|n} \mu(d)\,\frac{n}{d}$$

where $\mu(d) = 0$ if d is divisible by the square of some prime number, and, if d is not divisible by the square of any prime number, $\mu(d) = 1$ or -1 according as the number of (distinct) prime divisors of d is even or odd. This leads us to the function named after A. F. Möbius (1790-1868).

52.11 Definition: The function $\mu: \mathbb{N} \to \mathbb{Z}$, where

$\mu(1) = 1$,

$\mu(n) = (-1)^r$ if n is the product of r distinct prime numbers,

$\mu(n) = 0$ otherwise, i.e., if n is divisible by the square of a prime number,

is called the *Möbius function.*

For example, $\mu(1) = 1$, $\mu(2) = -1$, $\mu(3) = -1$, $\mu(4) = 0$, $\mu(5) = -1$, $\mu(6) = 1$, $\mu(7) = -1$, $\mu(8) = 0$, $\mu(9) = 0$, $\mu(10) = 1$.

The two formulas $n = \sum_{d|n} \varphi(d)$ and $\varphi(n) = \sum_{d|n} \mu(d)\frac{n}{d}$ are equivalent. This is a special case of a formula known as Möbius inversion formula that connects a divisor sum $\sum_{d|n} a_d$ with the a_d. To establish this formula, we need a lemma.

52.12 Lemma: *Let μ be the Möbius function and $n \in \mathbb{N}$. Then $\sum_{d|n} \mu(d)$ is equal to 1 in case $n = 1$ and to 0 in case $n > 1$.*

Proof: If $n = 1$, then $\sum_{d|n} \mu(d) = \sum_{d|1} \mu(d) = \mu(1) = 1$.

If $n > 1$ and $n = p_1^{a_1}p_2^{a_2}\dots p_k^{a_k}$ is the canonical decomposition of n into prime numbers, then

$$\begin{aligned}\sum_{d|n} \mu(d) = \sum_{d|p_1p_2\cdots p_k} \mu(d) &= \mu(1) + (\mu(p_1) + \mu(p_2) + \cdots + \mu(p_k)) \\ &\quad + (\mu(p_1p_2) + \mu(p_1p_3) + \cdots + \mu(p_{k-1}p_k)) \\ &\quad + (\mu(p_1p_2p_3) + \cdots + \mu(p_{k-2}p_{k-1}p_k)) \\ &\quad + \cdots \\ &\quad + \mu(p_1p_2\dots p_k) \\ &= 1 + \binom{k}{1}(-1)^1 + \binom{k}{2}(-1)^2 + \binom{k}{3}(-1)^3 + \cdots + \binom{k}{k}(-1)^k \\ &= (1-1)^k = 0. \end{aligned}$$

□

52.13 Lemma (Möbius inversion formula): *Let K be a field and let $f : \mathbb{N} \to K$ be any function. Define the function $F : \mathbb{N} \to K$ by declaring*

$$F(n) = \sum_{d|n} f(d).$$

for all $n \in \mathbb{N}$. Then $f(n) = \sum_{d|n} \mu(d)F(\frac{n}{d}) = \sum_{d|n} \mu(\frac{n}{d})F(d)$

for all $n \in \mathbb{N}$.

Proof: Let $n \in \mathbb{N}$. For any positive divisor d of n, we have

$$F(\frac{n}{d}) = \sum_{b|\frac{n}{d}} f(b),$$

$$\mu(d)F(\frac{n}{d}) = \sum_{b|\frac{n}{d}} \mu(d)f(b)$$

$$\sum_{d|n} \mu(d)F(\frac{n}{d}) = \sum_{d|n} \sum_{b|\frac{n}{d}} \mu(d)f(b).$$

The last sum is over all ordered pairs (d,b) of positive divisors of n such that $db|n$. Hence it is also the sum over all ordered pairs (b,d) of positive divisors of n such that $bd|n$ and we get

$$\sum_{d|n} \mu(d)F(\frac{n}{d}) = \sum_{b|n} \sum_{d|\frac{n}{b}} \mu(d)f(b) = \sum_{b|n} f(b)\Big(\sum_{d|\frac{n}{b}} \mu(d)\Big) = f(n)$$

since $\sum_{d|\frac{n}{b}} \mu(d)$ is equal to 1 when $b = n$ and to 0 when b is a proper divisor of n (Lemma 52.12). □

52.14 Lemma: *Let K be a field and let $f: \mathbb{N} \to K^\times$ be any function. Define the function $F: \mathbb{N} \to K^\times$ by declaring*

$$F(n) = \prod_{d|n} f(d).$$

for all $n \in \mathbb{N}$. Then $f(n) = \prod_{d|n} F(\frac{n}{d})^{\mu(d)} = \prod_{d|n} F(d)^{\mu(n/d)}$

for all $n \in \mathbb{N}$.

Proof: Let $n \in \mathbb{N}$. We have $F(\frac{n}{d}) = \prod_{b|\frac{n}{d}} f(b)$,

$$F(\frac{n}{d})^{\mu(d)} = \prod_{b|\frac{n}{d}} f(b)^{\mu(d)}$$

$$\prod_{d|n} F\left(\frac{n}{d}\right)^{\mu(d)} = \prod_{d|n} \prod_{b|\frac{n}{d}} f(b)^{\mu(d)}$$

and so

$$\prod_{d|n} F\left(\frac{n}{d}\right)^{\mu(d)} = \prod_{b|n} \prod_{d|\frac{n}{b}} f(b)^{\mu(d)} = \prod_{b|n} \left(f(b)^{\sum_{d|(n/b)} \mu(d)}\right) = f(n) \qquad \square$$

*

* *

We return to finite fields. We will prove that, for any prime number p and natural number n, there is a finite field with p^n elements and that any two finite fields with the same number of elements are isomorphic. We begin by discussing the decomposition of $x^{p^n} - x \in \mathbb{F}_p[x]$ into irreducible polynomials in the unique factorization domain $\mathbb{F}_p[x]$. It turns out that all irreducible factors of $x^{p^n} - x$ are distinct, and an irreducible polynomial in $\mathbb{F}_p[x]$ divides $x^{p^n} - x$ if and only if its degree divides n.

52.15 Theorem: *Let p be a positive prime number and let $F_d(x)$ be the product of all monic irreducible polynomials of degree d in $\mathbb{F}_p[x]$ (if there is no monic irreducible polynomial of degree d in $\mathbb{F}_p[x]$, let $F_d(x)$ be the constant polynomial $1 \in \mathbb{F}_p[x]$). Then*

$$x^{p^n} - x = \prod_{d|n} F_d(x) \qquad \text{in } \mathbb{F}_p[x].$$

Proof: All roots of $x^{p^n} - x$ are simple, because $x^{p^n} - x$ is relatively prime to its derivative -1. So $x^{p^n} - x$ is not divisible by the square of any polynomial in $\mathbb{F}_p[x]$. In particular, $x^{p^n} - x$ is not divisible by the square of any of its irreducible factors in $\mathbb{F}_p[x]$.

Suppose $f(x) \in \mathbb{F}_p[x]$ is a monic irreducible polynomial in $\mathbb{F}_p[x]$ and let $d = \deg f(x)$. We construct the field $\mathbb{F}_p(a)$ by adjoining a root a of $f(x)$ to $\mathbb{F}_p$. Now $f(x)$ is the minimal polynomial of a over $\mathbb{F}_p$, so $|\mathbb{F}_p(a):\mathbb{F}_p| = \deg f(x) = d$ and $\mathbb{F}_p(a)$ is a field of p^d elements. Therefore $b^{p^d} = b$ for all $b \in \mathbb{F}_p(a)$

(Lemma 52.4(3)). We are to prove that $f(x)|x^{p^n} - x$ in $\mathbb{F}_p[x]$ if and only if $d|n$ in $\mathbb{Z}$.

Assume $d|n$. As $a \in \mathbb{F}_p(a)$, we have $a^{p^d} = a$, so a is a root of $x^{p^d} - x \in \mathbb{F}_p[x]$. But $f(x)$ is the minimal polynomial of a over $\mathbb{F}_p$, hence $f(x)|x^{p^d} - x$ in $\mathbb{F}_p[x]$. From $d|n$, it follows that $x^{p^d} - x \,|x^{p^n} - x$ (Lemma 52.7(3)), so $f(x)|x^{p^n} - x$.

Assume now $f(x)|x^{p^n} - x$. Then $f(x)g(x) = x^{p^n} - x$ for some $g(x) \in \mathbb{F}_p[x]$, and $f(a)g(a) = a^{p^n} - a = 0$. So a is a root of $x^{p^n} - x$. But then any element of $\mathbb{F}_p(a)$ is a root of $x^{p^n} - x$: if $b \in \mathbb{F}_p(a)$, say $b = f_0 + f_1 a + f_2 a^2 + \ldots + f_{d-1}a^{d-1}$ with $f_0 f_1 f_2, \ldots f_{d-1} \in \mathbb{F}_p$, then we get

$$\begin{aligned} b^{p^n} &= (f_0 + f_1 a + f_2 a^2 + \cdots + f_{d-1}a^{d-1})^{p^n} \\ &= f_0^{p^n} + f_1^{p^n} a^{p^n} + f_2^{p^n} (a^2)^{p^n} + \cdots + (f_{d-1})^{p^n} (a^{d-1})^{p^n} \\ &= f_0 + f_1 a + f_2 a^2 + \cdots + f_{d-1}a^{d-1} = b. \end{aligned}$$

Since the elements of $\mathbb{F}_p(a)$ coincide with the roots of $x^{p^d} - x$ (Lemma 52.4(3)), we see that any root of $x^{p^d} - x$ is also a root of $x^{p^n} - x$. Therefore $x^{p^d} - x$ divides $x^{p^n} - x$ and, by Lemma 52.7(3), d divides n. □

52.16 Lemma: *Let p be a prime number and let N_d be the number of monic irreducible polynomials of degree d in $\mathbb{F}_p[x]$. Let $F_d(x)$ be the product of all the N_d monic irreducible polynomials of degree d in $\mathbb{F}_p[x]$ (with the understanding $F_d(x) = 1$ in case $N_d = 0$; we prove presently that $N_d > 0$). For any $n \in \mathbb{N}$, we have*

(1) $$p^n = \sum_{d|n} d\, N_d;$$

(2) $$F_n(x) = \prod_{d|n} (x^{p^d} - x)^{\mu(n/d)};$$

(3) $$N_n = \frac{1}{n} \sum_{d|n} \mu\left(\frac{n}{d}\right) p^d;$$

(4) $$N_n > 0.$$

Proof: (1) This follows from $x^{p^n} - x = \prod_{d|n} F_d(x)$ by equating the degrees of the polynomials on both sides.

(2) This follows from the same equation by Lemma 52.14 (with the function $F: \mathbb{N} \to \mathbb{F}_p(x)$ that maps $n \in \mathbb{N}$ to $F_n(x)$).

(3) This follows from part (1) by Möbius inversion formula (Lemma 52.13).

(4) $N_n \geqslant 0$ by its definition. Also, if $N_n = 0$, we get $\sum_{d|n} \mu(\frac{n}{d}) p^d = 0$ from part (3) and, dividing both sides by the smallest p^d for which $\mu(\frac{n}{d}) \neq 0$, say by p^{d_0}, we obtain an equation $-\mu(\frac{n}{d_0}) = \sum_{\substack{d|n \\ d \neq d_0}} \mu(\frac{n}{d}) p^{d-d_0}$, where the right hand side is and the left hand side is not divisible by p, a contradiction. Hence $N_n > 0$. □

52.17 Theorem: *Let $n \in \in \mathbb{N}$ and let p be a prime number. Then there exists a finite field with p^n elements.*

Proof: By Lemma 52.16(4), there is an irreducible polynomial $f(x)$ of degree n in $\mathbb{F}_p[x]$. Let K be the field obtained by adjoining a root of $f(x)$ to $\mathbb{F}_p$. Then $|K:\mathbb{F}_p| = n$ and K is a field with p^n elements (Theorem 50.7). □

52.18 Theorem: *Let K be a field and let G be a finite subgroup of $K^\times$. Then G is cyclic. In particular, if K is a finite field, then $K^\times$ is cyclic.*

Proof: Let $n = |G|$. The order of any element g in G is a divisor of n. Hence we have the disjoint union

$$G = \bigcup_{d|n} \{g \in G: o(g) = n\},$$

from which we obtain

$$n = |G| = \sum_{d|n} \psi(d),$$

where $\psi(d)$ is the number of elements in G of order d.

We claim that $\psi(d)$ is either 0 or $\varphi(d)$. If there is no element in G of order d, then of course $\psi(d) = 0$. If there does exist an element g in G of order

d, then all the d elements in the cyclic group $\langle g \rangle$ generated by g satisfy $g^d = 1$. Hence they are roots of the polynomial $x^d - 1 \in K[x]$ and this polynomial has therefore at least d roots in K. On the other hand, it can have at most d roots in K, thus it has exactly d roots in K, namely the elements in $\langle g \rangle$. Thus any element in G that has order d, which necessarily is a root of $x^d - 1$, is in the subgroup $\langle g \rangle$, and an element in $\langle g \rangle$ is of order d if and only if that element is a generator of $\langle g \rangle$. Thus the elements in G of order d coincide with the generators of $\langle g \rangle$. There are $\varphi(d)$ generators of $\langle g \rangle$, so there are $\varphi(d)$ elements in G of order d, i.e., $\psi(d) = \varphi(d)$, as claimed.

Since $\psi(d) \leqslant \varphi(d)$ for any positive divisor of n, we obtain $n = \sum_{d|n} \psi(d) \leqslant \sum_{d|n} \varphi(d) = n$ and this gives $\psi(d) = \varphi(d)$ for all positive divisors d of n. In particular, $\psi(n) = \varphi(n) > 0$: there is an element a in G of order n. Thus G is the cyclic group $\langle a \rangle$. □

52.19 Theorem: *Let K be a field of p^n elements and let t be a generator of the cyclic group K^*. Then*
(1) $K = \mathbb{F}_p(t)$.
(2) *The minimal polynomial of t over $\mathbb{F}_p$ has degree n.*
(3) *If K_1 is any field of p^n elements, then the minimal polynomial t over $\mathbb{F}_p$ has a root in K_1.*

Proof: Since $0 \in \mathbb{F}_p(t)$ and since any nonzero element of K, being a power of t, is in $\mathbb{F}_p(t)$, we get $K \subseteq \mathbb{F}_p(t)$; thus $K = \mathbb{F}_p(t)$. This proves (1). Then the degree of the minimal polynomial of t over $\mathbb{F}_p$ is equal to $|\mathbb{F}_p(t):\mathbb{F}_p| = |K:\mathbb{F}_p| = n$. This proves (2). Finally, since the degree of the minimal polynomial of t over $\mathbb{F}_p$ is equal to n, hence a divisor of n, this polynomial is a divisor of $x^{p^n} - x$ (Theorem 52.15) and has n distinct roots in K_1 (Lemma 52.4(3)); in particular, there is a root of this polynomial in K_1. This proves (3). □

52.20 Theorem: *Any two finite fields with the same number of elements are isomorphic.*

Proof: Let K and K_1 be fields of p^n elements. Then $K^\times$ is a cyclic group (Theorem 52.18). Let t be a generator of $K^\times$. Then $K = \mathbb{F}_p(t)$ by Theorem 52.19(1). Let $f(x) \in \mathbb{F}_p[x]$ be the minimal polynomial of t over $\mathbb{F}_p$. Now $f(x)$ has a root c in K_1 (Theorem 52.19(3)). Let $\mathbb{F}_p(c) \subseteq K_1$ be the subfield of K_1 generated by c over $\mathbb{F}_p$. Then $n = deg\ f(x) = |\mathbb{F}_p(c):\mathbb{F}_p| \leq |K_1:\mathbb{F}_p| = n$ yields $\mathbb{F}_p(c) = K_1$. We then get

$$K_1 = \mathbb{F}_p(c) \cong \mathbb{F}_p[x]/(f(x)) \cong \mathbb{F}_p(t) = K$$

from Theorem 50.6. Hence $K_1 \cong K$. □

In view of this theorem, we identify all finite fields of the same number of elements. Thus there is a unique field of q elements ($q = p^n$), and this field will be henceforward denoted by $\mathbb{F}_q$.

Exercises

1. Find finite subgroups of $\mathbb{C}^\times$ and show directly that they are cyclic.

2. Let E and K be finite fields, with $K \subseteq E$ and $|E:K| = 5$. Let $a \in K$. If there is no $b \in K$ such that $b^2 = a$, show that there is no $b \in E$ such that $b^2 = a$.

3. Let E and K be finite fields, with $K \subseteq E$ and let $|E:K| = n$. Let $a \in K$ be such that there is no $b \in K$ such that $b^2 = a$. Prove that, if n is odd, there is no $b \in E$ such that $b^2 = a$ and that, if n is even, there is a $b \in E$ such that $b^2 = a$.

4. Find all monic irreducible polynomials in $\mathbb{F}_2[x]$ of degree 2,3 and 4. Verify Lemma 52.16(3).

5. Let p and q be distinct prime numbers. Find the number of monic irreducible polynomials in $\mathbb{F}_p[x]$ of degree q.

6. Let K be a field with p^n elements. Let $a \in K$ and put $f(x) = \prod_{k=0}^{n-1} (x - a^{p^k})$. Show that $f(x) \in \mathbb{F}_p[x]$. Conclude that $a + a^p + a^{p^2} + \cdots + a^{p^{n-1}} \in \mathbb{F}_p$. This sum $a + a^p + a^{p^2} + \cdots + a^{p^{n-1}}$ is called the *trace of a over* $\mathbb{F}_p$ and is denoted by

$T_{K/\mathbb{F}_p}(a)$. Prove that $T_{K/\mathbb{F}_p}(a+b) = T_{K/\mathbb{F}_p}(a) + T_{K/\mathbb{F}_p}(b)$ and $T_{K/\mathbb{F}_p}(ca) = cT_{K/\mathbb{F}_p}(a)$ for all $a,b \in K$ and $c \in \mathbb{F}_p$ and show that there is an $a \in K$ with $T_{K/\mathbb{F}_p}(a) \neq 0$.

7. Keep the notation of Ex.6. Prove that $g(x) = x^p - x - a \in K[x]$ is either irreducible in $K[x]$ or is a product of p polynomials of degree one. Prove that the latter alternative holds if and only if $T_{K/\mathbb{F}_p}(a) = 0$.

8. Construct addition and multiplication tables for the finite fields $\mathbb{F}_4, \mathbb{F}_8, \mathbb{F}_9$ and $\mathbb{F}_{16}$.

9. Find a generator of the cyclic group $K^\times$ when $K = \mathbb{F}_4, \mathbb{F}_5, \mathbb{F}_7, \mathbb{F}_8, \mathbb{F}_9, \mathbb{F}_{16}, \mathbb{F}_{27}$.

10. Prove that a root of $x^2 + 7x + 2 \in \mathbb{F}_{11}[x]$ is a generator of $\mathbb{F}_{11^2}^\times$.

§53
Splitting Fields

Given a field K and a polynomial $f(x) \in K[x] \backslash K$, is it possible to find an extension field E of K such that $f(x)$ can be written as a product of polynomials in $E[x]$ of first degree? In this paragraph, we study this problem.

This problem is related to another important question in the theory of field extensions: whether a field isomorphism can be extended to a field isomorphism of the extension field. More precisely, if E_1/K_1 and E_2/K_2 are field extensions and if $\varphi: K_1 \to K_2$ is a field isomorphism, can we find a field isomorphism $\psi: E_1 \to E_2$ such that $\psi_{|K} = \varphi$? The answer is negative in general, but in the important case of simple algebraic extensions, it turns out to be positive.

Let us recall that, for any field isomorphism $\varphi: K_1 \to K_2$, we have a ring isomorphism $\hat{\varphi}: K_1[x] \to K_2[x]$ given by $\left(\sum_{i=0}^{m} a_i x^i\right)\hat{\varphi} = \sum_{i=0}^{m} (a_i\varphi)x^i$ (Lemma 33.7, Theorem 33.8).

53.1 Lemma: *Let E_1/K_1 and E_2/K_2 be field extensions and let $\varphi: K_1 \to K_2$ be a field isomorphism. Assume $f_1(x) \in K_1[x]$ is an irreducible polynomial in $K_1[x]$ and let $f_2(x) = (f_1(x))\hat{\varphi} \in K_2[x]$ be its image under $\hat{\varphi}$. Let $u_1 \in E_1$ be a root of $f_1(x)$ and $u_2 \in E_2$ a root of $f_2(x)$. Let $K_1(u_1) \subseteq E_1$ be the subfield of E_1 generated by u_1 and let $K_2(u_2) \subseteq E_2$ be the subfield of E_2 generated by u_2. Then φ extends to an isomorphism of fields $K_1(u_1) \cong K_2(u_2)$ that maps u_1 to u_2; that is, there is a field isomorphism $\psi: K_1(u_1) \to K_2(u_2)$ such that $u_1\psi = u_2$ and $\psi_{|K} = \varphi$. Moreover, there is only one isomorphism ψ with these properties.*

Proof: We make use of Theorem 50.6 and Theorem 30.18. Since u_1 is a root of $f_1(x)$ and $f_1(x)$ is irreducible in $K_1[x]$, we see that $c_0^{-1}f_1(x)$ is the minimal polynomial of u_1 over K_1, where c_0 is the leading coefficient of $f_1(x)$ (Theorem 50.3; as $f_1(x)$ is irreducible, it is not the zero polynomial or a polynomial of degree zero). Now $(c_0^{-1}f_1) = (f_1)$ and from Theorem 50.6 and its proof (which depends on Theorem 30.17 and Theorem 30.18), we know that

$$\alpha: K_1(u_1) \to K_1[x]/(f_1)$$

$$\sum_i a_i u_1^i \to \sum_i a_i (x + (f_1))^i$$

is a field isomorphism. Likewise there is a field isomorphism

$$\beta: K_2(u_2) \to K_2[x]/(f_2).$$

$$\sum_i a_i u_2^i \to \sum_i a_i (x + (f_2))^i$$

Besides, we have an isomorphism of rings

$$\hat{\varphi}: K_1[x] \to K_2[x].$$

Here (f_1) is an ideal of $K_1[x]$, therefore $Im\ \hat{\varphi}_{|(f_1)} = (f_2)$ is an ideal of $K_2[x]$ and $K_1[x]/(f_1) \cong K_2[x]/Im\ \hat{\varphi}_{|(f_1)} = K_2[x]/(f_2)$ by Theorem 30.19(7). More specifically, we have the isomorphism

$$\lambda: K_1[x]/(f_1) \to K_2[x]/(f_2).$$
$$g + (f_1) \to g\hat{\varphi} + (f_2)$$

Hence $\alpha\lambda\beta^{-1}: K_1(u_1) \to K_2(u_2)$ is a (ring, and therefore also a) field isomorphism. We write $\psi = \alpha\lambda\beta^{-1}$. Then $a\psi = (a\alpha)\lambda\beta^{-1} = a\lambda\beta^{-1} = [a + (f_1)]\lambda\beta^{-1} = [a + (f_2)]\beta^{-1} = a\beta^{-1} = a$ for any $a \in K_1$ (we regard K_1 as a subfield of $K_1[x]/(f_1)$ and K_2 as a subfield of $K_2[x]/(f_2)$ as in Kronecker's theorem (Theorem 51.1)) and $u_1\psi = (u_1\alpha)\lambda\beta^{-1} = (x + (f_1))\lambda\beta^{-1} = (x + (f_2))\beta^{-1} = u_2$. Thus ψ is an extension of φ such that $u_1\psi = u_2$.

The uniqueness of ψ as an extension of φ with $u_1\psi = u_2$ follows from the fact that powers of u_1 form a K_1-basis of $K_1(u_1)$ (Theorem 50.7). Indeed, if $\mu: K_1(u_1) \to K_2(u_2)$ is a field isomorphism such that $u_1\mu = u_2$ and $\mu_{|K} = \varphi$, then μ maps any element $t = \sum_i a_i u_1^i$ of $K_1(u_1)$, where $a_i \in K_1$, to $t\mu = \sum_i (a_i u_1^i)\mu = \sum_i a_i \mu (u_1\mu)^i = \sum_i a_i \varphi u_2^i = \left(\sum_i a_i u_1^i\right)\psi = t\psi$, and so $\mu = \psi$. □

53.2 Theorem: *Let E_1/K and E_2/K be field extensions and let $u_1 \in E_1$ and $u_2 \in E_2$ be algebraic over K. Then the minimal polynomial of u_1 over K coincides with the minimal polynomial of u_2 if and only if there is an isomorphism (necessarily unique) of fields $\psi: K(u_1) \to K(u_2)$ that maps u_1 to u_2 and whose restriction to K is the identity mapping on K.*

Proof: If u_1 and u_2 have the same minimal polynomial over K, then we apply Theorem 53.1 with the identity mapping $\iota: K \to K$ in place of φ and conclude that the identity isomorphism can be extended to a unique isomorphism $\psi: K(u_1) \to K(u_2)$ such that $u_1\psi = u_2$.

Conversely, suppose that $\psi: K(u_1) \to K(u_2)$ is a field isomorphism such that $u_1\psi = u_2$ and $a\psi = a$ for all $a \in K$. Let $f(x) = \sum_{i=0}^{m} a_i x^i$ be the minimal polynomial of u_1 over K. Then $0 = f(u_1) = \sum_{i=0}^{m} a_i u_1{}^i$. Hence $0 = 0\psi = \left(\sum_{i=0}^{m} a_i u_1{}^i\right)\psi$ $= \sum_{i=0}^{m} (a_i u_1{}^i)\psi = \sum_{i=0}^{m} a_i\psi u_1{}^i\psi = \sum_{i=0}^{m} a_i (u_1\psi)^i = \sum_{i=0}^{m} a_i u_2{}^i = f(u_2)$. Thus u_2 is a root of $f(x)$ and $f(x) \in K[x]$ is a monic irreducible polynomial, which means that $f(x)$ is the minimal polynomial of u_2 over K. □

53.3 Remark: Theorem 53.1 should not mislead the reader to believe that any field isomorphism can be extended to larger fields. Consider, for example, the isomorphism $\varphi: \mathbb{Q}(\sqrt{2}) \to \mathbb{Q}(\sqrt{2})$ given by $a + b\sqrt{2} \to a - b\sqrt{2}$ $(a,b \in \mathbb{Q})$. Now $\mathbb{Q}(\sqrt[4]{2})$ is an extension field of $\mathbb{Q}(\sqrt{2})$. If $\varphi: \mathbb{Q}(\sqrt{2}) \to \mathbb{Q}(\sqrt{2})$ could be extended to an isomorphism $\psi: \mathbb{Q}(\sqrt[4]{2}) \to \mathbb{Q}(\sqrt[4]{2})$, we would have $-\sqrt{2} = (\sqrt{2})\varphi = (\sqrt{2})\psi = ((\sqrt[4]{2})^2)\psi = ((\sqrt[4]{2})\psi)^2$, a contradiction, since the square of $(\sqrt[4]{2})\psi \in \mathbb{Q}(\sqrt[4]{2}) \subseteq \mathbb{R}$ has to be positive. So φ cannot be extended to an isomorphism of $\mathbb{Q}(\sqrt[4]{2})$.

The most important application of Theorem 53.1 is that any two splitting fields of a polynomial are isomorphic. We now discuss this matter.

53.4 Definition: Let E/K be a field extension and $f(x) \in K[x] \setminus K$. If $f(x)$ can be written as a product of linear polynomials in $E[x]$, i.e., if there are $a_0, a_1, a_2, \ldots, a_m$ in E such that $f(x) = a_0(x - a_1)(x - a_2)\ldots(x - a_m)$, then $f(x)$ is said to *split in* E. If $f(x)$ splits in E but not in any proper subfield of E containing K, then E is called a *splitting field of* $f(x)$ *over* K.

53.5 Examples: (a) Consider $x^2 + 1 \in \mathbb{R}[x]$. Now $x^2 + 1 = (x - i)(x + i)$ in $\mathbb{C}[x]$, so $x^2 + 1$ splits in $\mathbb{C}$. It does not split in any proper subfield of $\mathbb{C}$ containing $\mathbb{R}$ because $\mathbb{R}$ is the only proper subfield of $\mathbb{C}$ containing $\mathbb{R}$ and $x^2 + 1$ does not split in $\mathbb{R}[x]$. So $\mathbb{C}$ is a splitting field of $x^2 + 1$ over $\mathbb{R}$.

$\mathbb{C}$ is not a splitting field of $x^2 + 1$ over $\mathbb{Q}$, because $x^2 + 1$ splits in the field $\mathbb{Q}(i) \subset \mathbb{C}$. Now $x^2 + 1$ does not split in $\mathbb{Q}$ which is the only proper subfield of $\mathbb{Q}(i)$ containing $\mathbb{Q}$. Hence $\mathbb{Q}(i)$ is a splitting field of $x^2 + 1$ over $\mathbb{Q}$.

(b) $\mathbb{Q}(\sqrt{2})$ is a splitting field of $x^2 - 2 \in \mathbb{Q}[x]$ over $\mathbb{Q}$.

(c) $x^3 - 2 \in \mathbb{Q}[x]$ does not split in $\mathbb{Q}(\sqrt[3]{2})$ because $x^3 - 2$
$= (x - \sqrt[3]{2})(x^2 + \sqrt[3]{2}x + (\sqrt[3]{2})^2)$ in $\mathbb{Q}(\sqrt[3]{2})[x]$ and the second factor is irreducible in $\mathbb{Q}(\sqrt[3]{2})[x]$. On the other hand, $x^3 - 2 = (x - \sqrt[3]{2})(x - \omega\sqrt[3]{2})(x - \omega^2\sqrt[3]{2})$ in $\mathbb{Q}(\sqrt[3]{2}, \omega)[x]$, so $x^3 - 2$ splits in $\mathbb{Q}(\sqrt[3]{2}, \omega)[x]$. In fact $\mathbb{Q}(\sqrt[3]{2}, \omega)$ is a splitting field of $x^3 - 2$ over $\mathbb{Q}$. Notice that $\mathbb{Q}(\sqrt[3]{2}, \omega) = \mathbb{Q}(\sqrt[3]{2}, \omega\sqrt[3]{2}, \omega^2\sqrt[3]{2})$ is the field generated by the roots of $x^3 - 2$ over $\mathbb{Q}$.

(d) Let E/K be a field extension and $f(x) \in K[x]$ a polynomial of positive degree n. Assume that E contains n roots $a_1, a_2, \ldots, a_n$ of $f(x)$ (counted with multiplicity). Then $H = K(a_1, a_2, \ldots, a_n)$ is a splitting field of $f(x)$ over K. Indeed, with the leading coefficient $a_0 \in K$, we have the factorization $f(x)$ $= a_0(x - a_1)(x - a_2)\ldots(x - a_n)$ in $H[x]$ since each factor $x - a_k$ belongs to $H[x]$. Hence $f(x)$ splits in $H[x]$. On the other hand, if L is any intermediate field of E/K in which $f(x)$ splits, then $x - a_k$ is in $L[x]$ and so a_k is in L for all k, thus $\{a_1, a_2, \ldots, a_n\} \subseteq L$ and $H = K(a_1, a_2, \ldots, a_n) \subseteq L$. Hence $f(x)$ does not split in any proper subfield of H containing K. Therefore, H is a splitting field of $f(x)$ over K. This argument shows in fact that $K(a_1, a_2, \ldots, a_n)$ is the

unique intermediate field of E/K which is a splitting field of $f(x)$ over K. In particular, E is a splitting field of $f(x)$ if and only if $E = K(a_1, a_2, \ldots, a_n)$.

(e) Let E/K be a field extension, L an intermediate field of this extension and $f(x) \in K[x] \backslash K$. Assume that E is a spliting field of $f(x)$ over K. Then E is a spliting field of $f(x)$ over L, too, since $f(x)$ splits in E but not in any proper subfield of E containing K so that all the more so $f(x)$ does not split in any proper subfield of E containing L.

(f) Let p be prime. Any greatest common divisor of $x^{p^n} - x$ with its derivative $p^n x^{p^n-1} - 1 = -1$ is a unit in $\mathbb{F}_p[x]$. Hence $x^{p^n} - x \in \mathbb{F}_p[x]$ has no multiple roots (Theorem 35.18(2)). Thus an extension field of $\mathbb{F}_p$ in which $x^{p^n} - x$ splits must have at least the p^n distinct roots of $f(x)$. We know that $x^{p^n} - x$ splits in the field $\mathbb{F}_{p^n}$ with p^n elements (Lemma 52.4(3)). Thus $\mathbb{F}_{p^n}$ is a splitting field of $x^{p^n} - x$ over $\mathbb{F}_p$.

(g) Let E/K be a field extension and $f(x) \in K[x] \backslash K$. Let $a_1 \in E$ be a root of $f(x)$ and let $L = K(a_1)$ be the subfield of E generated by a_1 over K, so that $f(x) = (x - a_1)g(x)$ for some $g(x) \in L[x]$. We claim that, if ($g(x)$ has positive degree and) E is a splitting field of $g(x)$ over L, then E is also a splitting field of $f(x)$ over K. Indeed, if E is a splitting field of $g(x)$ over L, then $g(x) = c(x - a_2)\ldots(x - a_n)$ where $c \in K$ and $a_2, \ldots, a_n \in E$. We know that $E = L(a_2, \ldots, a_n)$ from Example 53.5(d). Then $f(x) = c(x - a_1)(x - a_2)\ldots(x - a_n)$ in $E[x]$ and $f(x)$ splits in $E[x]$. On the other hand, if E' is any intermediate field of E/K and $f(x)$ splits in E', then $c(x - a_1)(x - a_2)\ldots(x - a_n)$ in $E'[x]$, so $a_1 \in E'$, so $L = K(a_1) \subseteq E'$ and $a_2, \ldots, a_n \in E'$, so $L(a_2, \ldots, a_n) \subseteq E'$ and $E \subseteq E'$. Thus $f(x)$ cannot split in any proper subfield of E containing K and E is a splitting field of $f(x)$ over K.

(h) We saw in Example 53.5(c) that $\mathbb{Q}(\sqrt[3]{2}, \omega)$ is a splitting field of $x^3 - 2$ over $\mathbb{Q}$. Likewise, $\mathbb{Q}(\sqrt[3]{2})[y]/(y^2 + y + 1)$ and $\mathbb{Q}(\omega)[y]/(y^3 - 2)$ are splitting fields of $x^3 - 2$ over $\mathbb{Q}$ (here y is an indeterminate over $\mathbb{Q}$). In these fields, $x^3 - 2$ splits as

$$[x - (\sqrt[3]{2} + (y^2 + y + 1))][(x - (\sqrt[3]{2}y + (y^2 + y + 1))][x - (\sqrt[3]{2}y^2 + (y^2 + y + 1))]$$

and $[x - (y + (y^3 - 2))][x - (\omega y + (y^3 - 2))][x - (\omega^2 y + (y^3 - 2))]$, respectively.

A natural question is whether any polynomial has a splitting field. We show now this is indeed the case. The following theorem is due to Kronecker.

53.6 Theorem: *Let $f(x)$ be an arbitrary polynomial of positive degree over an arbitrary field K. Then there is an extension field E of K such that $|E:K| \leqslant (\deg f(x))!$ and E is a splitting field of $f(x)$ over K.*

Proof: We make induction on $n = \deg f(x)$. If $n = 1$, then $f(x) = c(x - a)$ for some $c,a \in K$ and so K is a splitting field of $f(x)$ over K and we have $|E:K| = 1 \leqslant 1!$. So the claim is established when $n = 1$.

Suppose now $\deg f(x) = n \geqslant 2$ and the theorem is true for any polynomial over any field if its degree is $n - 1$. We construct an extension field L of K in which $f(x)$ has a root a and $|L:K| \leqslant n$ (Theorem 51.5; possibly $L = K$). Then, by theorem 35.6, $f(x) = (x - a)g(x)$ for some $g(x)$ in $L[x]$. Now $\deg g(x) = n - 1$ and, by induction, there is an extension field E of L such that E is a splitting field of $g(x)$ over L and $|E:L| \leqslant (n - 1)!$. From Example 53.5(g), we conclude that E is a splitting field of $f(x)$ over K. Moreover, $|E:K| = |E:L||L:K| \leqslant (n - 1)!|L:K| \leqslant (n - 1)!n = n!$. □

We see that Theorem 53.6 is nothing but repeated application of Kronecker's theorem (Theorem 51.5). We use Theorem 51.5 succesively until we find a field which contains *all* the roots of $f(x)$. In the proof of Theorem 53.6, the successive adjunction of roots is replaced by an inductive argument.

We now turn to the question of uniqueness. Example 53.5(h) reveals that there can be many distinct splitting fields of a polynomial. However, as has already been remarked, all splitting fields of a polynomial are isomorphic. We prove a slightly more general theorem.

53.7 Theorem: *Let E_1/K_1 and E_2/K_2 be field extensions and let $\varphi: K_1 \to K_2$ be a field isomorphism. Let $f_1(x) \in K_1[x]$ be a polynomial in $K_1[x] \setminus K_1$ and let $f_2(x) = (f_1(x))\hat{\varphi} \in K_2[x] \setminus K_2$ be its image under $\hat{\varphi}$. If E_1 is a splitting field*

of $f_1(x)$ over K_1 and E_2 is a splitting field of $f_2(x)$ over K_2, then φ extends to a field isomorphism $\Phi: E_1 \to E_2$ and so $E_1 \cong E_2$.

Proof: E_1 is generated over K_1 by the roots of $f_1(x)$. Since each root of $f_1(x)$ is algebraic over K_1 and there are finitely many roots, Theorem 50.12 yields that $|E_1:K_1|$ is finite. We make induction on $|E_1:K_1|$.

If $|E_1:K_1| = 1$, then $E_1 = K_1$ and $f_1(x)$ splits in K_1. Then $f_2(x)$ splits in K_2 and $K_2 = E_2$. Thus $E_1 = K_1 \xrightarrow{\varphi} K_2 = E_2$ is the desired isomorphism.

Suppose now $|E_1:K_1| \geqslant 2$ and suppose that any field isomorphism can be extended to an isomorphism of splitting fields of corresponding polynomials whenever the degree of a splitting field is less than or equal to $n - 1$. Since $|E_1:K_1| \geqslant 2$ and E is generated over K_1 by the roots of $f_1(x)$, there must be a root of $f_1(x)$ in E_1 which does not belong to K_1. Let u_1 be a root of $f_1(x)$ in $E_1 \backslash K_1$. Assume $g_1(x) \in K_1[x]$ is the minimal polynomial of u_1 over K_1 and let u_2 be a root of $(g_1(x))\hat{\varphi} = g_2(x) \in K_2[x]$ in E_2. From Lemma 53.1, we know that φ can be extended to an isomorphism $\psi: K_1(u_1) \to K_2(u_2)$. Now $u_1 \in E_1 \backslash K_1$, so $|K_1(u_1):K_1| > 1$ and $|E_1:K_1(u_1)| < n$ (Theorem 48.13). As E_1 is a splitting field of $f_1(x)$ over $K_1(u_1)$ and E_2 is a splitting field of $f_2(x)$ over $K_2(u_2)$ (Example 53.5(e)), we conclude, by induction, that ψ can be extended to an isomorphism $\Phi: E_1 \to E_2$. This Φ is the desired extension of φ. □

53.8 Theorem: *Let K be a field and let $f(x)$ be any polynomial of positive degree in $K[x]$. Then any two splitting fields of $f(x)$ over K are isomorphic. In fact, the splitting fields of $f(x)$ are isomorphic by an isomorphism fixing each element of K.*

Proof: Let E_1 and E_2 be splitting fields of $f(x)$ over K and apply Theorem 53.7 with $K_1 = K = K_2$ and $\varphi = \iota$ = identity mapping on K. □

In the remainder of this paragraph, we discuss algebraically closed fields.

53.9 Definition: A field K is said to be *algebraically closed* if K has no proper algebraic extension field, i.e., if any algebraic extension E of K coincides with K.

53.10 Theorem: *Let K be a field. The following statements are equivalent.*

(i) *K is algebraically closed.*

(ii) *Any irreducible polynomial in $K[x]$ has degree one.*

(iii) *Every polynomial of positive degree in $K[x]$ has a root in K.*

(iv) *Every polynomial of positive degree in $K[x]$ splits in K.*

Proof: (i) $\Rightarrow$ (ii) Assume that K is algebraically closed. If there were an irreducible polynomial $f(x)$ in $K[x]$ with $deg\ f(x) > 1$, then $E = K[x]/(f)$ would be an algebraic extension of K with $K \subset E$, contrary to the assumption that K has no proper algebraic extension. Thus every irreducible polynomial in $K[x]$ has degree one.

(ii) $\Rightarrow$ (i) Suppose that any irreducible polynomial in $K[x]$ has degree one. We want to show that K has no proper algebraic extension. If E were a proper algebraic extension of K, then there would be an $a \in E \backslash K$. Now a is algebraic over K and $K \subset K(a)$ since $a \notin K$ (Lemma 49.6(1)). This leads to the contradiction

$$1 < |K(a):K| = \text{degree of the minimal polynomial of } a \text{ over } K$$
$$= \text{degree of an irreducible polynomial in } K[x] = 1.$$

Thus K is algebraically closed.

(ii) $\Rightarrow$ (iii) Suppose that any irreducible polynomial in $K[x]$ has degree one. Let $f(x)$ be any polynomial of positive degree in $K[x]$. We show that $f(x)$ has a root in K. Indeed, any irreducible divisor of $f(x)$ has the form $c(x - a)$ with $c, a \in K$ and thus has a root a in K, so $f(x)$ too, has a root in a in K.

(iii) $\Rightarrow$ (iv) Assume that any polynomial of positive degree in $K[x]$ has a root in K and let $f_1(x) \in K[x] \backslash K$. Then $f_1(x)$ has a root a_1 in K and $f_1(x) = (x - a_1)f_2(x)$ for some $f_2(x) \in K[x]$. If $f_2(x)$ has positive degree, then $f_2(x)$ has a root a_2 in K and $f_2(x) = (x - a_2)f_3(x)$ for some $f_3(x) \in K[x]$; so $f_1(x) = (x - a_1)(x - a_2)f_3(x)$. If $f_3(x)$ has positive degree, then $f_3(x)$ has a root a_3 in K and $f_3(x) = (x - a_3)f_4(x)$ for some $f_4(x) \in K[x]$; so $f_1(x)$

$= (x - a_1)(x - a_2)(x - a_3)f_4(x)$. Proceeding in this way, we will meet an $f_n(x)$ of degree zero and $f_1(x) = (x - a_1)(x - a_2)(x - a_3)\ldots(x - a_n)f_n$ splits in K.

(iv) $\Longrightarrow$ (ii) Suppose every polynomial of positive degree in $K[x]$ splits in K and let $f(x)$ be an irreducible polynomial in $K[x]$. Then, by assumption, $f(x)$ is a product of $deg\ f(x)$ polynomials of degree one. Since $f(x)$ is irreducible, the number $deg\ f(x)$ of factors must be one: $deg\ f(x) = 1$. So any irreducible polynomial in $K[x]$ has degree one. □

An example of an algebraically closed field is $\mathbb{C}$. This is a consequence of the result known as the fundamental theorem of algebra, which says that any polynomial with complex coefficients has a root in $\mathbb{C}$. The name 'fundamental theorem of algebra' is grotesque, for this is neither a fundamental theorem nor a theorem of algebra! Any proof of this result is bound to use some results from analysis.

53.11 Lemma: *Let E/K be a field extension and assume that E is algebraically closed. Let A be the algebraic closure of K in E (Definition 50.15). Then A is an algebraically closed field.*

Proof: It suffices to prove that any polynomial in $A[x]\setminus A$ has a root in A. Let $f(x)$ be a polynomial of positive degree in $A[x]$. Then $f(x)$ is a polynomial of positive degree in $E[x]$ and therefore has a root b in E (Theorem 53.10). Then $A(b)$ is an algebraic extension of A and A is an algebraic extension of K, so $A(b)$ is an algebraic extension of K (Theorem 50.16). Consequently $b \in A(b)$ is algebraic over K and hence $b \in A$ by the definition of A. □

53.12 Definition: Let E/K be a field extension. If E is an algebraic extension of K and E is algebraically closed, then E is called an *algebraic closure of K*.

Does every field K have an algebraic closure? The answer is 'yes' and its proof requires Zorn's Lemma. There is no algebraic difficulty in the

proof, but there are certain set-theoretical subtelties and we will not give the proof in this book. It is also true that an algebraic closure of a field K is unique in the sense that any two algebraic closures of a field K are isomorphic by an isomorphism that fixes each element of K.

Exercises

1. Construct a splitting field over $\mathbb{Q}$ of
   (a) $x^2 - 3$;
   (b) $x^2 - 5$;
   (c) $x^2 - p$, where $p \in \mathbb{N}$ is a prime number;
   (d) $x^5 - 1$;
   (e) $x^p - 1$, where $p \in \mathbb{N}$ is a prime number;
   (f) $x^4 - 5x^2 + 6$;
   (g) $x^6 - 10x^4 + 31x^2 - 30$;
   (h) $x^5 + 3x^4 + x^3 - 8x^2 - 6x + 4$;
   (i) $x^4 - x^2 + 1$.

2. Let K be a field and let $f(x) \in K[x]$ be of degree $n > 0$. If E is a splitting field of $f(x)$ over K, show that $[E:K]$ divides $n!$.

3. What is the difference between an algebraic closure of a field K and the algebraic closure of K in an extension field?

4. Prove that a finite field cannot be algebraically closed.

§54
Galois Theory

This paragraph gives an exposition of Galois theory. Given any field extension E/K, we associate intermediate fields of E/K with subgroups of a group, called the Galois group of the extension. Many questions about the intermediate field structure of the extension can be thus reduced to related questions about the subgroup structure of the Galois group. Our exposition closely follows the treatment of I. Kaplansky.

If E/K is a field extension, then E is a field and also a K-vector space. It will be very fruitful to study both the field and the vector space structure of E at the same time. For this reason, we consider mappings which preserve both of these structures.

Let E be a field. Let us recall that a field automorphism φ of E is a one-to-one ring homomorphism from E onto E. Equivalently, a field automorphism of E is an automorphism of the additive group $(E,+)$ which is also a ring isomorphism of E. Clearly the identity mapping on E is a field automorphism of E, so the set of all field automorphisms of E is not empty. Besides, if φ and ψ are any two field automorphisms of E, then $\varphi\psi$ and φ^{-1} are automorphisms of the additive group $(E,+)$ which are ring isomorphisms from E onto E as well (Lemma 30.16); thus $\varphi\psi$ and φ^{-1} are field automorphisms of E. Therefore the set of all field automorphisms of E is a subgroup of the group of all automorphisms of the additive group $(E,+)$. The group of all field automorphisms of E will be denoted by $Aut(E)$. Thus we use the same notation for the group of additive group automorphisms of E and the group of field automorphisms of E. This is not likely to cause confusion. Anyhow, $Aut(E)$ will play a minor role in the sequel.

$Aut(E)$ is the collection of mappings from E onto E that preserve the field structure of E. From these field automorphisms, we select the mappings that preserve the vector space structure of E. We introduce some terminology.

54.1 Definition: Let E/K and F/K be field extensions. A mapping $\varphi: E \to F$ is called a *K-homomorphism* if φ is both a field homomorphism and a K-vector space homomorphism. A K-homomorphism $\varphi: E \to F$ is called a *K-isomorphism* if φ is one-to-one and onto F. A K-isomorphism from E onto E is called a *K-automorphism of E*. The set of all K-automorphisms of E will be denoted by $Aut_K E$ or by $G(E/K)$.

If $\varphi: E \to F$ is a K-homomorphism, then $(1_E)\varphi = 1_F$ (see the remarks following Definition 48.9) since φ is a field homomorphism and, for any $k \in K$, there holds $k\varphi = (k1_E)\varphi = k(1_E)\varphi = k1_F = k$ since φ is a K-linear transformation. Thus $k\varphi = k$ for all $k \in K$. Conversely, if $\varphi: E \to F$ is a K-homomorphism such that $k\varphi = k$ for all $k \in K$, then $(ke)\varphi = k\varphi \cdot e\varphi = k(e\varphi)$ for all $k \in K$ and $e \in E$, and thus φ is a K-linear transformation, too. Therefore a field homomorphism $\varphi: E \to F$ is a K-homomorphism if and only if φ fixes every element of K.

54.2 Lemma: *Let E/K be a field extension and let $Aut_K E$ be the set of all K-automorphisms of E over K. Then $Aut_K E$ is a group.*

Proof: We have $\iota_E \in Aut_K E \subseteq Aut(E)$ and $Aut(E)$ is a group. Since the composition two vector space isomorphisms and also the inverse of a vector space isomorphism are vector space isomorphisms (Theorem 41.10), $Aut_K E$ is closed under composition and forming of inverses. Thus $Aut_K E$ is a subgroup of $Aut(E)$. □

54.3 Definition: Let E/K be a field extension. The group $Aut_K E = G(E/K)$ is called the *Galois group of E over K*.

54.4 Examples: (a) Let E be any field and let P be the prime subfield of E. Any field automorphism φ of E fixes $1 \in E$. This implies that φ fixes each element in P. Therefore any field automorphism of E is a P-automorphism of E and $Aut(E) = Aut_P(E)$.

(b) The familiar complex conjugation mapping ($a + bi \to a - bi$, where $a,b \in \mathbb{R}$) is an $\mathbb{R}$-automorphism of $\mathbb{C}$.

(c) The mapping $\varphi: \mathbb{Q}(\sqrt{2}) \to \mathbb{Q}(\sqrt{2})$ that maps $a + b\sqrt{2}$ to $a - b\sqrt{2}$ (where $a,b \in \mathbb{Q}$) is a $\mathbb{Q}$-automorphism of $\mathbb{Q}(\sqrt{2})$.

(d) Let K be a field and x an indeterminate over K. Then $K(x)$ is an extension field of K. If $a \in K^*$, then ax is transcendental over K and, by Theorem 49.10, the mapping $\sigma_a: K(x) \to K(x)$ given by $f(x)/g(x) \to f(ax)/g(ax)$ is a field automorphism of $K(x)$. It is easy to see that σ_a is in fact a K-automorphism of $K(x)$. Likewise, for any $b \in K$, the mapping τ_b: $K(x) \to K(x)$ given by $f(x)/g(x) \to f(x + b)/g(x + b)$ is a K-automorphism of $K(x)$. As $x\sigma_a\tau_b = (ax)\tau_b = a(x + b) \neq ax + b = (x + b)\sigma_a = x\tau_b\sigma_a$ unless $a \neq 1$ or $b \neq 0$, we see that $Aut_K K(x)$ is a nonabelian group.

We find $Aut_K K(x)$. In the following, y and z are two additional distinct indeterminates over K.

Let u be an arbitrary element in $K(x)\backslash K$, say $u = p(x)/q(x)$, where $p(x)$ and $q(x)$ are relatively prime polynomials in $K[x]$ and $q(x) \neq 0$. We claim that u is transcendental over K and $K(x)$ is finite dimensional (hence algebraic) over $K(u)$.

We prove the first claim, viz. that u is transcendental over K. If u were algebraic over K, then u would have a minimal polynomial

$$H(y) = y^k + c_{k-1}y^{k-1} + \cdots + c_1 y + c_0 \in K[y]$$

over K. Then, from $H(u) = 0$, we would get

$$(p(x)/q(x))^k + c_{k-1}(p(x)/q(x))^{k-1} + \cdots + c_1(p(x)/q(x)) + c_0 = 0,$$

$$p(x)^k + c_{k-1}p(x)^{k-1}q(x) + \cdots + c_1 p(x)q(x)^{k-1} + c_0 q(x)^k = 0,$$

$$q(x) | p(x)^k \text{ in } K[x] \text{ and } (p(x),q(x)) \approx 1,$$

$$q(x) \text{ is a unit in } K[x], \text{ so } q(x) \in K,$$

$$u = p(x)/q(x) \in K[x],$$

$H(u) = u^k + c_{k-1}u^{k-1} + \cdots + c_1 u + c_0$ is a polynomial of degree $k(deg.\ p(x))$, contrary to $H(u) = 0$. Thus u is transcendental over K.

Secondly, we prove that $|K(x):K(u)|$ is finite. Now $u = p(x)/q(x)$. Let us put $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, $q(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$, with $a_n \neq 0 \neq b_m$. We note that x is a root of the polynomial

$$F(y) = (b_m u)y^m + (b_{m-1}u)y^{m-1} + \cdots + (b_1 u)y + b_0$$
$$- a_n y^n - a_{n-1}y^{n-1} - \cdots - a_1 y - a_0$$

in $K(u)[y]$. Thus x is algebraic over $K(u)$. We see moreover that $deg\ F(y) = max\ (m,n) = max\ (deg\ p(x),deg\ q(x))$, because $b_m u - a_n \neq 0$ as $u \notin K$. We will show that $F(y)$ is irreducible over $K(u)$. This will imply $cF(y)$ is the minimal polynomial of x over $K(u)$, where $1/c$ is the leading coefficient of $F(y)$, and so $|K(x):K(u)| = deg\ cF(y) = deg\ F(y) = max\ (deg\ p(x),deg\ q(x))$.

Now the irreducibility of $F(y)$ over $K(u)$. Since u is transcendental over K, the substitution homomorphism $z \to u$ is in fact a field isomorphism from $K(z)$ onto $K(u) \subseteq K(x)$ (Theorem 49.10). So $K(u) \cong K(z)$ and Theorem 33.8 gives $K(u)[y] \cong K(z)[y]$. Then $F(y) \in K(u)[y]$ is irreducible in $K(u)[y]$ if and only if its image $F(z) \in K(z)[y]$ is irreducible in $K(z)[y]$. From Theorem 34.5(3) and Lemma 34.11, we conclude $F(z)$ is irreducible in $K(z)[y]$ if and only if $F(z) = q(y)z - p(y)$ is irreducible in $K[z][y] = K[y][z]$. But $F(z) = q(y)z - p(y)$ is certainly irreducible in $K[y][z]$ since $q(y)z - p(y)$ is of degree one in $K[y][z]$ and its coefficients $q(y)$, $-p(y)$ are relatively prime in $K[y]$ (for $p(x)$ and $q(x)$ are relatively prime in $K[x]$).

Thus we get $|K(x):K(u)| = max\ (deg\ p(x),deg\ q(x))$ for any $u = p(x)/q(x)$ in $K(x)\backslash K$, where $p(x)$ and $q(x)$ are relatively prime polynomials in $K[x]$ and $q(x) \neq 0$.

Now let $\varphi \in Aut_K K(x)$ and $x\varphi = u$. Write $u = p(x)/q(x)$ as above. Since

$$\begin{aligned} K(u) = K(x\varphi) &= \{f(x\varphi)/g(x\varphi): f,g \in K[x],\ g \neq 0\} \\ &= \{f(x)\varphi/g(x)\varphi: f,g \in K[x],\ g \neq 0\} \\ &= \{(f(x)/g(x))\varphi: f,g \in K[x],\ g \neq 0\} \\ &= (K(x))\varphi = K(x) \neq K, \end{aligned}$$

we have $u \in K(x)\backslash K$ and

$$1 = |K(x):K(x)| = |K(x):K(u)| = max\ (deg\ p(x),deg\ q(x))$$

yields $p(x) = ax + b$, $q(x) = cx + d$ for some $a,b,c,d \in K$. Here $ad - bc \neq 0$ for $ad - bc = 0$ implies the contradiction $u = p(x)/q(x) = (ax + b)/(cx + d) \in K$.

Thus every automorphism in $Aut_K K(x)$ is a substitution homomorphism that sends x to $(ax + b)/(cx + d)$ for some $a,b,c,d \in K$ satisfying $ad - bc \neq 0$. Conversely, if φ is a substitution homomorphism of this type, with $x\varphi = (ax + b)/(cx + d)$, $a,b,c,d \in K$, $ad - bc \neq 0$, then $(ax + b)/(cx + d) =: u$ is not in K, so u is transcendental over K and φ is a field homomorphism from $K(x)$ onto $K(u)$. Since $ad - bc \neq 0$, both of a and c cannot be 0, so $|K(x):K(u)| = max\ (deg\ ax + b,deg\ cx + d) = 1$ and $K(u) = K(x)$. Hence φ is a field homomorphism from $K(x)$ onto $K(x)$. As φ fixes all elements in K, we infer that φ is in $Aut_K K(x)$. Therefore $Aut_K K(x)$ consists exactly of the

substitution homomorphisms $x \to (ax + b)/(cx + d)$, where $a,b,c,d \in K$ and $ad - bc \neq 0$.

The next lemma is a generalization of the familiar fact that the complex conjugate of any root of a polynomial with real coefficients is also a root of the same polynomial. In the terminology of §26, if E/K is a field extension, $Aut_K E$ acts on the set of distinct roots of a polynomial $f(x)$ over K.

54.5 **Lemma:** *Let E/K be a field extension and $f(x) \in K[x]$. If $u \in E$ is a root of $f(x)$, then, for any $\varphi \in Aut_K E$, the element $u\varphi$ of E is also a root of $f(x)$.*

Proof: If $f(x) = \sum_{i=0}^{m} a_i x^i$, then $f(u) = 0$ implies $0 = 0\varphi = (f(u))\varphi = (\sum_{i=0}^{m} a_i u^i)\varphi$ $= \sum_{i=0}^{m} (a_i\varphi)(u^i\varphi) = \sum_{i=0}^{m} a_i(u\varphi)^i = f(u\varphi)$. Thus $u\varphi$ is a root of $f(x)$. □

Let E/K be a finite dimensional extension and assume that $\{a_1,a_2, \ldots,a_m\}$ is a K-basis of E. Then any K-automorphism of E is completely determined by its effect on the basis elements, for if φ and ψ are K-automorphisms of E and $a_i\varphi = a_i\psi$ for $i = 1,2, \ldots, m$, then, for any $a \in E$, which we write in the form $\sum_{i=0}^{m} k_i a_i$, we have $a\varphi = (\sum_{i=0}^{m} k_i a_i)\varphi = \sum_{i=0}^{m} k_i\varphi a_i\varphi$ $= \sum_{i=0}^{m} k_i(a_i\varphi) = \sum_{i=0}^{m} k_i(a_i\psi) = \sum_{i=0}^{m} k_i\psi a_i\psi = (\sum_{i=0}^{m} k_i a_i)\psi = a\psi$. For this reason, we will describe the K-automorphisms of E by describing the images of the basis elements. Thus the conjugation mapping will be denoted by $i \to -i$, the mapping of Example 54.4(c) by $\sqrt{2} \to -\sqrt{2}$, etc.

In particular, if E/K is a simple extension and a is a primitive element, then $\{1,a,a^2, \ldots,a^{n-1}\}$ is a K-basis of $E = K(a)$, where n is the degree of the minimal polynomial of a over K (Theorem 50.7). Let $\varphi \in Aut_K E$. Since $a^i\varphi = (a\varphi)^i$ for any $i = 0,1,2, \ldots,n - 1$, the mapping φ is completely determined

by its effect on a. Now $a\varphi$ is a root in $K(a)$ of the minimal polynomial of a over K. Thus $|Aut_K E| \leq r$, where r is the number of distinct roots in $K(a)$ of the minimal polynomial of a over K. We proved the following lemma.

54.6 Lemma: *Let K be a field. If a is algebraic over K with the minimal polynomial f over K, and if r is the number of distinct roots of f in $K(a)$ then $|Aut_K K(a)| \leq r \leq deg\ f = |K(a):K|$.* □

54.7 Examples: (a) Let $\sqrt[3]{2}$ be the positive real cube root of 2. Thus $\mathbb{Q}(\sqrt[3]{2}) \subseteq \mathbb{R}$. We find $Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt[3]{2})$. If $\varphi \in Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt[3]{2})$, then $\sqrt[3]{2}\varphi \in \mathbb{R}$ is a root of the minimal polynomial $x^3 - 2$ of $\sqrt[3]{2}$ over $\mathbb{Q}$. Since the roots of $x^3 - 2$ other than $\sqrt[3]{2}$ are complex, $\sqrt[3]{2}\varphi$ must be $\sqrt[3]{2}$. Thus φ must be the identity mapping on $\mathbb{Q}(\sqrt[3]{2})$ and $Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt[3]{2}) = 1$.

(b) $\mathbb{C} = \mathbb{R}(i)$ and the minimal polynomial of i over $\mathbb{R}$ is $x^2 + 1$, which has two roots in $\mathbb{C}$. Thus $|Aut_{\mathbb{R}}\mathbb{C}| \leq 2$. Since the identity mapping and conjugation mapping are $\mathbb{R}$-automorphisms of $\mathbb{C}$, $|Aut_{\mathbb{R}}\mathbb{C}| = 2$ and we get $Aut_{\mathbb{R}}\mathbb{C} \cong C_2$. Likewise $Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt{2}) \cong C_2$.

(c) We find $Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt{2},\sqrt{3})$. We have $\mathbb{Q}(\sqrt{2},\sqrt{3}) = \mathbb{Q}(\sqrt{2})(\sqrt{3})$. Here $\{1,\sqrt{2}\}$ is a $\mathbb{Q}$-basis of $\mathbb{Q}(\sqrt{2})$ and $\{1,\sqrt{3}\}$ is a $\mathbb{Q}(\sqrt{2})$-basis of $\mathbb{Q}(\sqrt{2})(\sqrt{3})$ (because $x^2 - 3$ is irreducible over $\mathbb{Q}(\sqrt{2})$), hence, by Theorem 48.13, $\{1,\sqrt{2},\sqrt{3},\sqrt{6}\}$ is a $\mathbb{Q}$-basis of $\mathbb{Q}(\sqrt{2},\sqrt{3})$. Now any $\varphi \in Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt{2},\sqrt{3})$ maps $\sqrt{2}$ to $\sqrt{2}$ or to $-\sqrt{2}$ and $\sqrt{3}$ to $\sqrt{3}$ or to $-\sqrt{3}$ and there are four possibilities for φ:

$$\begin{aligned}
(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6})\varphi_1 &= a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \\
(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6})\varphi_2 &= a + b\sqrt{2} - c\sqrt{3} - d\sqrt{6} \\
(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6})\varphi_3 &= a - b\sqrt{2} + c\sqrt{3} - d\sqrt{6} \\
(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6})\varphi_4 &= a - b\sqrt{2} - c\sqrt{3} + d\sqrt{6}
\end{aligned}$$

$(a,b,c,d \in \mathbb{Q})$. It is easy to see that $\varphi_1,\varphi_2,\varphi_3,\varphi_4$ are indeed $\mathbb{Q}$-automorphisms of $\mathbb{Q}(\sqrt{2},\sqrt{3})$ so that $Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt{2},\sqrt{3}) = \{\varphi_1,\varphi_2,\varphi_3,\varphi_4\}$. Here φ_1 is the

identity mapping on $\mathbb{Q}(\sqrt{2},\sqrt{3})$ and $\varphi_i^2 = \varphi_1$, $\varphi_i\varphi_j = \varphi_k$ when $\{i,j,k\} = \{2,3,4\}$. Thus $Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt{2},\sqrt{3}) \cong C_2 \times C_2 \cong V_4$.

We now proceed to establish the correspondence between intermediate fields of an extension E/K and subgroups of $Aut_K E$.

54.8 Lemma: *Let E/K be a field extension and put $G = Aut_K E$.*

(1) *If L is an intermediate field of E/K, then*

$$L' = \{\varphi \in G: l\varphi = l \text{ for all } l \in L\}$$

is a subgroup of G.

(2) *If H is a subgroup of G, then*

$$H' = \{a \in E: a\varphi = a \text{ for all } \varphi \in H\}$$

is an intermediate field of E/K.

Proof: (1) Clearly $\iota_E \in L'$, so $L' \neq \varnothing$. If $\varphi,\psi \in L'$, then $l(\varphi\psi) = (l\varphi)\psi = l\psi = l$ for all $l \in L$, so $\varphi\psi \in L'$ and $l\varphi = l$ gives $l = l\varphi^{-1}$ for all $l \in L$, so $\varphi^{-1} \in L'$. Thus L' is a subgroup of $Aut_K E$. (In fact $L' = Aut_L E$.)

(2) Since any $\varphi \in H' \subseteq Aut_K E$ fixes the elements of K, we have $K \subseteq H'$. If $a,b \in H'$, then $a\varphi = a$ and $b\varphi = b$ for all $\varphi \in H$, so

$$\begin{aligned} (a+b)\varphi = a\varphi + b\varphi = a + b, &\qquad a + b \in H', \\ (-b)\varphi = -(b\varphi) = -b, &\qquad -b \in H', \\ (ab)\varphi = a\varphi \cdot b\varphi = ab, &\qquad ab \in H', \\ (1/b)\varphi = 1/b\varphi = 1/b \quad (\text{provided } b \neq 0), &\qquad 1/b \in H'. \end{aligned}$$

So H' is a subfield of E and therefore H' is an intermediate field of E/K. □

For example, in the notation of Example 54.7(c), we have

$$\mathbb{Q}(\sqrt{2},\sqrt{3})' = 1 \leqslant G = Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt{2},\sqrt{3})$$

$$\mathbb{Q}(\sqrt{2})' = \{\varphi_1,\varphi_2\}, \qquad \mathbb{Q}(\sqrt{3}) = \{\varphi_1,\varphi_3\}, \qquad \mathbb{Q}(\sqrt{6})' = \{\varphi_1,\varphi_4\},$$

$$\mathbb{Q}' = G$$

and $1' = \mathbb{Q}(\sqrt{2},\sqrt{3})$

$$\{\varphi_1,\varphi_2\}' = \mathbb{Q}(\sqrt{2}), \qquad \{\varphi_1,\varphi_3\}' = \mathbb{Q}(\sqrt{3}), \qquad \{\varphi_1,\varphi_4\}' = \mathbb{Q}(\sqrt{6}),$$

$$G' = \mathbb{Q}.$$

If E/K is a field extension and $H \leqslant Aut_K E$, then H' is called the *fixed field of* H. Let us consider the four extreme cases of the priming correspondence in Lemma 54.8.

54.9 Lemma: *Let E/K be a field extension and $G = Aut_K E$. Then*

(1) $1' = E$.

(2) $E' = 1$.

(3) $K' = G$.

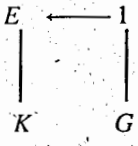
(4) G' *contains* K, *and possibly* $K \subset G'$.

Proof: (1) $1' = \{a \in E: a\varphi = a \text{ for all } \varphi \in 1\} = \{a \in E: a\iota_E = a\} = E$.

(2) $E' = \{\varphi \in G: a\varphi = a \text{ for all } a \in E\} = \{\iota_E\} = 1$.

(3) $K' = \{\varphi \in G: a\varphi = a \text{ for all } a \in K\} = G$.

(4) Of course $K \subseteq G'$. From Example 54.7(a), we know that $Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt[3]{2}) = 1$ so that, for the extension $\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}$, we have $G = 1$ and $K = \mathbb{Q} \subset \mathbb{Q}(\sqrt[3]{2}) = 1'$ $= G'$. Thus G' is not always equal to K. □



54.10 Definition: Let E/K be a field extension and put $G = Aut_K E$. If G' is equal to K, then E/K is said to be a *Galois extension* and E is said to be *Galois over* K.

Equivalently, E/K is Galois if and only if for any element a of $E\backslash K$, there exists a $\varphi \in Aut_K E$ such that $a\varphi \neq a$. It is easy to verify that $\mathbb{C}$ is a Galois extension of $\mathbb{R}$ and that $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{2},\sqrt{3})$ are Galois extensions of $\mathbb{Q}$.

54.11 Lemma: *Let E/K be a field extension and put $G = Aut_K E$. Let L,M be intermediate fields of E/K and let H,J be subgroups of G. If X is an intermediate field of E/K or a subgroup of G, we denote $(X')'$ shortly by X''. Then the following hold.*

(1) *If $L \subseteq M$, then $M' \leqslant L'$.*
(2) *If $H \leqslant J$, then $J' \subseteq H'$.*
(3) *$L \subseteq L''$ and $H \leqslant H''$.*
(4) *$E''' = L'$ and $H''' = H'$.*

Proof: (1) Suppose $L \subseteq M$. If $\varphi \in M'$, then $a\varphi = a$ for all $a \in M$ and a fortiori $a\varphi = a$ for all $a \in L$, hence $\varphi \in L'$ and consequently $M' \subseteq L'$.

(2) Suppose $H \leqslant J$. If $a \in J'$, then $a\varphi = a$ for all $\varphi \in J$ and a fortiori $a\varphi = a$ for all $\varphi \in H$, hence $a \in H'$ and consequently $J' \subseteq H'$.

(3) If $a \in L$, then $a\varphi = a$ for all $\varphi \in L'$ by the definition of L', so a is fixed by all the K-automorphisms in L'. Hence a is in the fixed field of L' and $a \in L''$. This gives $L \subseteq L''$. If $\varphi \in H$, then $a\varphi = a$ for all $a \in H'$ by the definition of H', so φ fixes every element in H', so $\varphi \in H''$. This gives $H \subseteq H''$.

(4) By parts (1) and (2), priming reverses inclusion, therefore $L \subseteq L''$ and $H \leqslant H''$ yield $L''' \leqslant L'$ and $H''' \subseteq H'$. Also, using (3) with L replaced by H' and H by L', we get $H' \subseteq H'''$ and $L' \leqslant L'''$. So $L''' = L'$ and $H''' = H'$. □

In general, L may very well be a proper subset of L'' and H a proper subset of H''. We introduce a term for the case of equality.

54.12 Definition: Let E/K be a field extension and $G = Aut_K E$. An intermediate field L of E/K is said to be *closed* if $L = L''$ and a subgroup H of G is said to be *closed* if $H = H''$.

So E is Galois over K if and only if K is closed. Lemma 54.11(4) states that any primed object is closed.

54.13 Theorem: *Let E/K be a field extension and $G = Aut_K E$. There is a one-to-one correspondence between the set of all closed intermediate fields of E/K and the set of all closed subgroups of G, given by $L \to L'$.*

Proof: If L is a closed intermediate field of E/K, then L' is a subgroup of G by Lemma 54.8(1) and L' is closed by Lemma 54.11(4). Thus priming is a mapping from the set of all closed intermediate fields of the extension into the set of all closed subgroups of G. This mapping is one-to-one, for $L' = M'$ implies $(L')'' = (M')''$, whence $L = M$ by Lemma 54.11(4) again. Finally, the priming mapping is *onto* the set of all closed subgroups of G because, if H is any closed subgroup of G, then H' is a closed intermediate field and $(H')' = H$. This completes the proof. □

This theorem is "virtually useless" until we determine which intermediate fields and which subgroups are closed. In the most important case when E/K is a finite dimensional Galois extension, all intermediate fields and all subgroups will turn out to be closed.

Our next goal is to show that an object is closed if it is "bigger than a closed object by a finite amount" (Theorem 54.16). We need two technical lemmas.

If E/K is a field extension and L,M are intermediate fields with $L \subseteq M$, then the dimension $|M:L|$ of M over L will be called the *relative dimension of L and M*. If G is the Galois group of this extension and H,J are subgroups of G with $H \leqslant J$, then the index $|J:H|$ of H in J will be called the *relative index of H and J*.

54.14 Lemma: *Let E/K be a field extension and L,M intermediate fields with $L \subseteq M$. If the relative dimension $|M:L|$ of L and M is finite, then the*

relative index of M' and L' is also finite. In fact, $|L':M'| \leqslant |M:L|$. In particular, if E/K is a finite dimensional extension, then $|Aut_K E| \leqslant |E:K|$.

Proof: We make induction on $n = |M:L|$. If $n = 1$, then $M = L$ and $L' = M'$, so $|L':M'| = 1$. Suppose now $n \geqslant 2$ and that the theorem has been proved for all $i < n$. Since $|M:L| > 1$, we can find an $a \in M \setminus L$. Now $|M:L|$ is finite and therefore M is an algebraic extension of L (Theorem 50.10), so a is algebraic over L. Let $f(x) \in L[x]$ be the minimal polynomial of a over L and put $k = deg\ f(x)$. We have $k > 1$ because $a \notin L$ (Lemma 49.6(1)). From Theorem 50.7, we deduce $|L(a):L| = k$ and Theorem 48.13 gives $|M:L(a)| = n/k$. The situation is depicted below.

$$\begin{array}{cccc} & M & \longrightarrow & M' \\ n/k & \cup & & \| \\ & L(a) & \longrightarrow & L(a)' \\ k & \cup & & \| \\ & L & \longrightarrow & L' \end{array}$$

In case $k < n$, induction settles everything: from $n/k < n$ and $k < n$, we obtain $|L(a)':M'| \leqslant |M:L(a)|$ and $|L':L(a)'| \leqslant |L(a):L|$ and therefore $|L':M'| = |L':L(a)'|\,|L(a)':M'| \leqslant |L(a):L|\,|M:L(a)| = k(n/k) = n = |M:L|$. The case $k = n$ requires a separate argument.

Suppose now $k = n$ so that $|M:L(a)| = 1$ and $M = L(a)$. In order to prove $|L':M'| \leqslant n$, we construct a one-to-one mapping from the set $\mathfrak{R}$ of all right cosets of M' in L' into the set of all distinct roots of $f(x)$. Since $\mathfrak{R}$ has $|L':M'|$ right cosets, this will prove that $|L':M'| \leqslant r$, where r is the number of distinct roots of $f(x)$ in M. As $r \leqslant deg\ f = |L(a):L| = |M:L|$, the theorem will be thereby proved.

What the required mapping should be is suggested by Lemma 54.5. We put

$$\begin{array}{rcl} \alpha: \mathfrak{R} & \to & \{b \in M: f(b) = 0\} \\ M'\varphi & \to & a\varphi \end{array}$$

($\varphi \in L'$). Since a is a root of $f(x)$ and $\varphi \in L' \leqslant G = Aut_K E$, Lemma 54.5 yields that $a\varphi$ is indeed a root of $f(x)$. The mapping α is well defined, for if $M'\varphi = M'\psi$ ($\varphi, \psi \in L'$), then $\varphi = \mu\psi$ for some $\mu \in M'$, so μ fixes every element of M, so μ fixes a and $(M'\varphi)\alpha = a\varphi = a(\mu\psi) = (a\mu)\psi = a\psi = (M'\psi)\alpha$. Moreover, α is one-to-one, for if $(M'\varphi)\alpha = (M'\psi)\alpha$, then $a\varphi = a\psi$, so $a\varphi\psi^{-1} = a$, so $\varphi\psi^{-1}$ fixes a, so $\varphi\psi^{-1}$ fixes each element of $L(a) = M$, so $\varphi\psi^{-1} \in M'$ and $M'\varphi = M'\psi$. This completes the proof of $|L':M'| \leqslant |M:L|$.

The assertion $|Aut_K E| \leqslant |E:K|$ follows easily: $|Aut_K E| = |Aut_K E:1| = |K':1| = |K':E'| \leqslant |E:K|$. □

54.15 Lemma: *Let E/K be a field extension and H,J are subgroups of $G = Aut_K E$ with $H \leqslant J$. If the relative index $|J:H|$ of H and J is finite, then the relative dimension of J' and H' is also finite. In fact, $|H':J'| \leqslant |J:H|$.*

Proof: Let $|J:H| = n$ and assume, by way of contradiction, that $|H':J'| > n$. Then there are $n + 1$ elements in H' that are linearly independent over J', say $a_1, a_2, \ldots, a_n, a_{n+1}$. Let $\bigcup_{i=1}^{n} H\varphi_i$ be the disjoint decomposition of J as a union of right cosets of H.

We consider the system of n linear equations in $n + 1$ unknowns:

$$\begin{aligned}
(a_1\varphi_1)x_1 + (a_2\varphi_1)x_2 + (a_3\varphi_1)x_3 + \cdots + (a_{n+1}\varphi_1)x_{n+1} &= 0 \\
(a_1\varphi_2)x_1 + (a_2\varphi_2)x_2 + (a_3\varphi_2)x_3 + \cdots + (a_{n+1}\varphi_2)x_{n+1} &= 0 \\
\cdots\cdots\cdots\cdots\cdots\cdots & \qquad (b)' \\
(a_1\varphi_n)x_1 + (a_2\varphi_n)x_2 + (a_3\varphi_n)x_3 + \cdots + (a_{n+1}\varphi_n)x_{n+1} &= 0
\end{aligned}$$

where the coefficients $a_i\varphi_j$ are in the field E. Since the number of unknowns is greater than the number of equations, this system (b) has a nontrivial solution in E (Theorem 45.1). From the nontrivial solutions of (b), we choose one for which the number of zeroes among x_i is as small as possible. Let $x_1 = b_1, x_2 = b_2, x_3 = b_3, \ldots, x_{n+1} = b_{n+1}$ be such a solution. Assume r of the b_j are nonzero ($r \leqslant n + 1$). By the choice of b_j, there is no solution $x_1 = c_1, x_2 = c_2, x_3 = c_3, \ldots, x_{n+1} = c_{n+1}$ of (b) in which the number of nonzero c_j's is less than r.

Eventually after renumbering, we may assume that $b_1, \ldots, b_r$ are distinct from zero and (in case $n + 1 > r$) $b_{r+1} = \cdots = b_{n+1} = 0$. Also, we may assume that $b_1 = 1$, for otherwise we may take the solution $b_1/b_1, b_2/b_1, b_3/b_1, \ldots, b_{n+1}/b_1$ instead of $b_1, b_2, b_3, \ldots, b_{n+1}$. Of course the number r of nonzero elements in both solutions are the same.

Let $\psi \in J$. We consider the system:

$$(a_1\varphi_1\psi)x_1 + (a_2\varphi_1\psi)x_2 + (a_3\varphi_1\psi)x_3 + \cdots + (a_{n+1}\varphi_1\psi)x_{n+1} = 0$$
$$(a_1\varphi_2\psi)x_1 + (a_2\varphi_2\psi)x_2 + (a_3\varphi_2\psi)x_3 + \cdots + (a_{n+1}\varphi_2\psi)x_{n+1} = 0$$
$$\cdots\cdots\cdots\cdots\cdots\cdots \qquad \text{(s)}$$
$$(a_1\varphi_n\psi)x_1 + (a_2\varphi_n\psi)x_2 + (a_3\varphi_n\psi)x_3 + \cdots + (a_{n+1}\varphi_n\psi)x_{n+1} = 0$$

We make two remarks concerning (s). First, since $x_1 = b_1 = 1, x_2 = b_2, x_3 = b_3, \ldots, x_{n+1} = b_{n+1}$ is a solution of (b) and ψ is a homomorphism, it is clear that $x_1 = b_1\psi = 1, x_2 = b_2\psi, x_3 = b_3\psi, \ldots, x_{n+1} = b_{n+1}\psi$ is a solution of (s). Second, the system (s) is identical with (b), aside from the order of the equations. To prove the last assertion, we note that $\varphi_1\psi, \varphi_2\psi, \varphi_3\psi, \ldots, \varphi_n\psi$ are elements of distinct right cosets of H in J, for $H\varphi_i\psi = H\varphi_j\psi$ implies $(\varphi_i\psi)(\varphi_j\psi)^{-1} \in H$, so $\varphi_i\varphi_j^{-1} \in H$, so $H\varphi_i = H\varphi_j$, so $i = j$. Let us write then

$$H\varphi_1\psi = H\varphi_{i_1},\ H\varphi_2\psi = H\varphi_{i_2},\ H\varphi_3\psi = H\varphi_{i_3}, \ldots, H\varphi_n\psi = H\varphi_{i_n}$$

so that $\quad \varphi_1\psi = \eta_1\varphi_{i_1},\ \varphi_2\psi = \eta_2\varphi_{i_2},\ \varphi_3\psi = \eta_3\varphi_{i_3}, \ldots, \varphi_n\psi = \eta_n\varphi_{i_n}$

for some $\eta_1, \eta_2, \eta_3, \ldots, \eta_n \in H$ (where $\{i_1, i_2, i_3, \ldots, i_n\} = \{1, 2, \ldots, n\}$). Thus each η_k fixes each a_m in H' and the i_k-th equation

$$(a_1\varphi_{i_k})x_1 + (a_2\varphi_{i_k})x_2 + (a_3\varphi_{i_k})x_3 + \cdots + (a_{n+1}\varphi_{i_k})x_{n+1} = 0$$

in (b) is identical with

$$(a_1\eta_k\varphi_{i_k})x_1 + (a_2\eta_k\varphi_{i_k})x_2 + (a_3\eta_k\varphi_{i_k})x_3 + \cdots + (a_{n+1}\eta_k\varphi_{i_k})x_{n+1} = 0$$

and therefore with the k-th equation

$$(a_1\varphi_k\psi)x_1 + (a_2\varphi_k\psi)x_2 + (a_3\varphi_k\psi)x_3 + \cdots + (a_{n+1}\varphi_k\psi)x_{n+1} = 0$$

in (s). This proves that (b) and (s) are identical systems.

Consequently, the solution $x_1 = 1, x_2 = b_2\psi, x_3 = b_3\psi, \ldots, x_{n+1} = b_{n+1}\psi$ of (s) is also a solution of (b). Now $x_1 = 1, x_2 = b_2, x_3 = b_3, \ldots, x_{n+1} = b_{n+1}$ is a solution of (b). Hence the difference of these solutions

$$x_1 = 0,\ x_2 = b_2 - b_2\psi,\ x_3 = b_3 - b_3\psi, \ldots, x_{n+1} = b_{n+1} - b_{n+1}\psi$$

i.e., $\quad x_1 = 0,\ x_2 = b_2 - b_2\psi, \ldots, x_r = b_r - b_r\psi, x_{r+1} = 0, \ldots, x_{n+1} = 0 \quad$ (c)

is a solution of (b).

So far, ψ was an arbitrary element of J. We now make a judicious choice of ψ. One of the $\varphi_1, \varphi_2, \varphi_3, \ldots, \varphi_n$ belongs to H, say $\varphi_1 \in H$, so $a_m\varphi_1 = a_m$ because $a_m \in H'$ for $m = 1, 2, \ldots, n, n+1$. Since $x_1 = b_1, x_2 = b_2, x_3 = b_3, \ldots, x_{n+1} = b_{n+1}$ is a solution of (b), we get

$$a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_{n+1}b_{n+1} = 0$$

from the first equation in (b). Here $\{a_1,a_2,a_3 \ldots,a_{n+1}\}$ is linearly independent over J' and $b_1 = 1 \neq 0$. Thus all of $b_1,b_2,b_3 \ldots,b_{n+1}$ cannot be in J': one of them, say b_2, is not in J'. So there is a $\psi \in J$ such that $b_2\psi \neq b_2$.

We choose $\psi \in J$ such that $b_2\psi \neq b_2$. Then the solution (c) of the system (b) is a nontrivial solution in which the number of zonzero elements is less than r, contrary to the meaning of r as the smallest number of nonzero elements in any solution of (b). This contradiction shows that $|H':J'| > n$ is impossible. Hence $|H':J'| \leq n = |J:H|$. □

54.16 Theorem: *Let E/K be a field extension and $G = Aut_K E$. Let L,M be intermediate fields of E/K with $L \subseteq M$ and let H,J be subgroups of G with $H \leq J$.*
(1) *If L is closed and $|M:L|$ is finite, then M is closed and $|L':M'| = |M:L|$.*
(2) *If H is closed and $|J:H|$ is finite, then J is closed and $|H':J'| = |J:H|$.*

Proof: (1) Here $M \subseteq M''$ by Lemma 54.11(3) and $L = L''$ by hypothesis, so
$$|M:L| \leq |M'':M|\,|M:L| = |M'':L| = |M'':L''| = |(M')':(L')'| \leq |L':M'| \leq |M:L|,$$
the last two inequalities by Lemma 54.15 and Lemma 54.14, respectively. This proves $|L':M'| = |M:L|$. The proof of (2) is similar and will be omitted. □

We are now in a position to state and prove the major theorem of this paragraph.

54.17 Theorem (Fundamental theorem of Galois theory): *Let E/K be a finite dimensional Galois extension of fields and $G = Aut_K E$. Then there is a one-to-one correspondence between the set of all intermediate fields of E/K and the set of all subgroups of G, given by $L \to L'$. In this correspondence, the relative dimension of two intermediate fields is equal to the relative index of the corresponding subgroups. In particular, $|G| = |Aut_K E| = |E:K|$.*

Proof: By Theorem 54.13, there is a one-to-one correspondence between the set of all closed intermediate fields of E/K and the set of all closed subgroups of G, given by $L \to L'$. Now K is closed (E/K is a Galois exten-

sion) by hypothesis and all intermediate fields are closed by Theorem 54.16(1) since they are finite dimensional over K. Moreover, if M is any intermediate field, then $|K':M'| = |M:K|$. In particular, E is closed and $|Aut_K E| = |G| = |G:1| = |G:E'| = |K':E'| = |E:K|$. Hence G is finite. Since 1 is closed, it follows from Theorem 54.16(2) that all subgroups of G are closed, because they are finite subgroups of G. Hence the priming mapping is a one-to-one correspondence between the set of *all* intermediate fields of E/K and the set of *all* subgroups of G. Theorem 54.16 tells that the relative dimension $|M:L|$ of two intermediate fields $L \subseteq M$ is equal to the relative index $|L':M'|$ of the corresponding subgroups of G and that the relative index $|J:H|$ of two subgroups $H \leqslant J$ of G is equal to the relative dimension $|H':J'|$ of the corresponding intermediate fields. □

54.18 Examples: (a) Let $\sqrt[3]{2}$ be the real cube root of 2 and consider the extension $\mathbb{Q}(\sqrt[3]{2},\omega)$ over $\mathbb{Q}$. The $\mathbb{Q}$-automorphisms of $\mathbb{Q}(\sqrt[3]{2},\omega)$ are $\varphi_1,\varphi_2,\varphi_3$, $\varphi_4,\varphi_5,\varphi_6$, where

$$\begin{aligned}
\varphi_1&: \sqrt[3]{2} \to \sqrt[3]{2}, && \omega \to \omega,\\
\varphi_2&: \sqrt[3]{2} \mapsto \sqrt[3]{2}, && \omega \to \omega^2 = -1 - \omega,\\
\varphi_3&: \sqrt[3]{2} \to \sqrt[3]{2}\omega, && \omega \to \omega,\\
\varphi_4&: \sqrt[3]{2} \to \sqrt[3]{2}\omega, && \omega \to \omega^2 = -1 - \omega,\\
\varphi_5&: \sqrt[3]{2} \to \sqrt[3]{2}\omega^2 = \sqrt[3]{2}(-1-\omega), && \omega \to \omega,\\
\varphi_6&: \sqrt[3]{2} \to \sqrt[3]{2}\omega^2 = \sqrt[3]{2}(-1-\omega), && \omega \to \omega^2 = -1 - \omega.
\end{aligned}$$

Any element u of $\mathbb{Q}(\sqrt[3]{2},\omega)$ can be written uniquely in the form

$$u = a + b\sqrt[3]{2} + c\sqrt[3]{4} + d\omega + e\sqrt[3]{2}\omega + f\sqrt[3]{4}\omega,$$

where a,b,c,d,e,f are rational numbers. We show that $\mathbb{Q}(\sqrt[3]{2},\omega)$ is Galois over $\mathbb{Q}$. To this end, we have to show that the fixed field of G is exactly $\mathbb{Q}$. Since

$$(a + b\sqrt[3]{2} + c\sqrt[3]{4} + d\omega + e\sqrt[3]{2}\omega + f\sqrt[3]{4}\omega)\varphi_2$$
$$= a + b\sqrt[3]{2} + c\sqrt[3]{4} + (d + e\sqrt[3]{2} + f\sqrt[3]{4})\omega^2$$
$$= a + b\sqrt[3]{2} + c\sqrt[3]{4} + (d + e\sqrt[3]{2} + f\sqrt[3]{4})(-1 - \omega)$$
$$= (a - d) + (b - e)\sqrt[3]{2} + (c - f)\sqrt[3]{4} - d\omega - e\sqrt[3]{2}\omega - f\sqrt[3]{4}\omega,$$

we see that an element $u = a + b\sqrt[3]{2} + c\sqrt[3]{4} + d\omega + e\sqrt[3]{2}\omega + f\sqrt[3]{4}\omega$ of $\mathbb{Q}(\sqrt[3]{2},\omega)$ is fixed by φ_2 if and only if

$$\begin{aligned} a &= a - d, \quad & d &= -d \\ b &= b - e, & e &= -e \\ c &= c - f, & f &= -f. \end{aligned}$$

So an element u of $\mathbb{Q}(\sqrt[3]{2},\omega)$ fixed by φ_2 has the form $a + b\sqrt[3]{2} + c\sqrt[3]{4}$. If u is fixed also by φ_3, then $a + b\sqrt[3]{2} + c\sqrt[3]{4} = (a + b\sqrt[3]{2} + c\sqrt[3]{4})\varphi$

$$= a + b\sqrt[3]{2}\omega + c\sqrt[3]{4}\omega^2$$
$$= a + b\sqrt[3]{2}\omega + c\sqrt[3]{4}(-1 - \omega)$$
$$= a - c\sqrt[3]{4} + b\sqrt[3]{2}\omega - c\sqrt[3]{4}\omega$$

yields $b = 0$, $c = -c$, $-c = 0$ and so $u = a \in \mathbb{Q}$. Since an element u in the fixed field of G is necessarily fixed by φ_2 and φ_3, that u has to be rational. Thus the fixed field of G is $\mathbb{Q}$. This shows that $\mathbb{Q}(\sqrt[3]{2},\omega)$ is Galois over $\mathbb{Q}$.

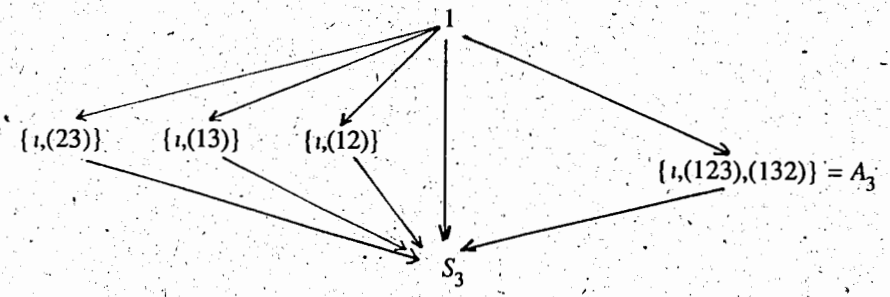
The multiplication table of $G(\mathbb{Q}(\sqrt[3]{2},\omega)/\mathbb{Q})$ can be constructed easily. Since $\sqrt[3]{2}\varphi_2\varphi_3 = \sqrt[3]{2}\varphi_3 = \sqrt[3]{2}\omega$ and $\omega\varphi_2\varphi_3 = \omega^2\varphi_3 = \omega^2$, we have $\varphi_2\varphi_3 = \varphi_4$ etc. and the multiplication table of $G(\mathbb{Q}(\sqrt[3]{2},\omega)/\mathbb{Q})$ is

	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
φ_1	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6
φ_2	φ_2	φ_1	φ_4	φ_3	φ_6	φ_5
φ_3	φ_3	φ_6	φ_5	φ_2	φ_1	φ_4
φ_4	φ_4	φ_5	φ_6	φ_1	φ_2	φ_3
φ_5	φ_5	φ_4	φ_1	φ_6	φ_3	φ_2
φ_6	φ_6	φ_3	φ_2	φ_5	φ_4	φ_1

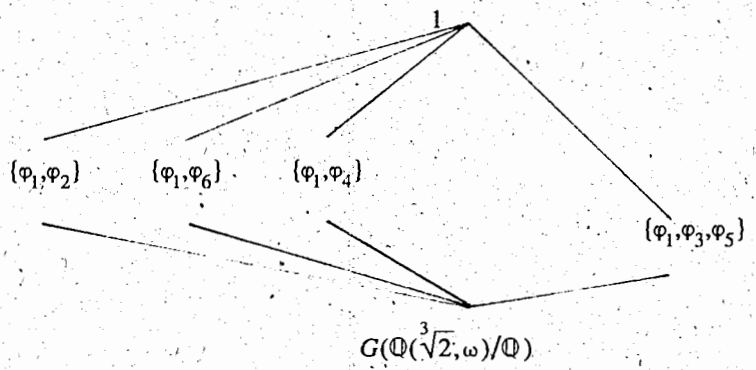
So $G(\mathbb{Q}(\sqrt[3]{2},\omega)/\mathbb{Q})$ is a nonabelian group of order 6 and isomorphic to S_3, as can be easily seen by comparing the table above with the multiplication table of S_3:

	ι	(23)	(123)	(12)	(132)	(13)
ι	ι	(23)	(123)	(12)	(132)	(13)
(23)	(23)	ι	(12)	(123)	(13)	(132)
(123)	(123)	(13)	(132)	(23)	ι	(12)
(12)	(12)	(132)	(13)	ι	(23)	(123)
(132)	(132)	(12)	ι	(13)	(123)	(23)
(13)	(13)	(123)	(23)	(132)	(12)	ι

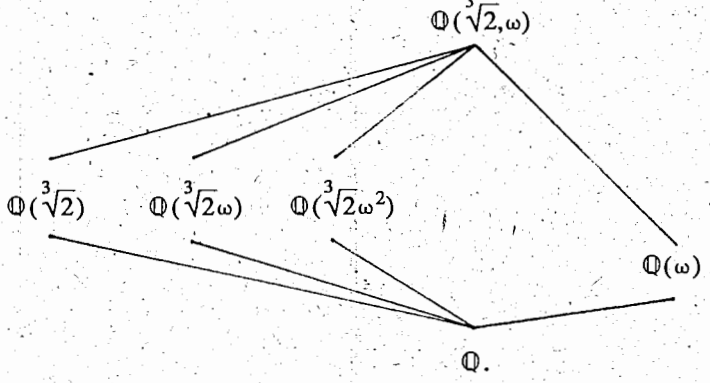
The isomorphism $G(\mathbb{Q}(\sqrt[3]{2},\omega)/\mathbb{Q}) \cong S_3$ can be found in a better way by observing that any automorphism in$G(\mathbb{Q}(\sqrt[3]{2},\omega)/\mathbb{Q})$ is completely determined by its effect on the roots of $x^3 - 2$. The roots of $x^3 - 2$ are $u_1 = \sqrt[3]{2}$, $u_2 = \sqrt[3]{2}\omega$, $u_3 = \sqrt[3]{2}\omega^2$. Now φ_2 maps u_1 to u_1, u_2 to u_3 and u_3 to u_2 and can therefore be represented, in a readily understood extension of the notation for permutations, as $\begin{pmatrix} u_1 & u_2 & u_3 \\ u_1 & u_3 & u_2 \end{pmatrix} = (u_1)(u_2u_3) = (u_2u_3)$. Dropping u and retaining only the indices, we see that φ_2 can be thought of as the permutation (23) in S_3. The other φ_j can be thought of as permutations in S_3 in a similar way and this gives the isomorphism $G(\mathbb{Q}(\sqrt[3]{2},\omega)/\mathbb{Q}) \cong S_3$. In the multiplication tables above, φ_j and its image in S_3 under this isomorphism occupy corresponding places. The subgroup structure of S_3 is well known and is depicted below ($A \to B$ means $A \subseteq B$).



So the subgroups of $G(\mathbb{Q}(\sqrt[3]{2},\omega)/\mathbb{Q})$ are



and priming yields



(b) Let $\sqrt[4]{2}$ be the real fourth root of 2 and consider the extension $\mathbb{Q}(\sqrt[4]{2},i)$ over $\mathbb{Q}$. The $\mathbb{Q}$-automorphisms of $\mathbb{Q}(\sqrt[4]{2},i)$ are $\varphi_1,\varphi_2,\varphi_3,\varphi_4,\varphi_5,\varphi_6,\varphi_7,\varphi_8$ where

$$\varphi_1: \sqrt[4]{2} \to \sqrt[4]{2}, \qquad i \to i,$$

$$\varphi_2: \sqrt[4]{2} \to \sqrt[4]{2}, \qquad i \to -i,$$

$$\varphi_3: \sqrt[4]{2} \to \sqrt[4]{2}i, \qquad i \to i,$$

$$\varphi_4: \sqrt[4]{2} \to \sqrt[4]{2}i, \qquad i \to -i,$$

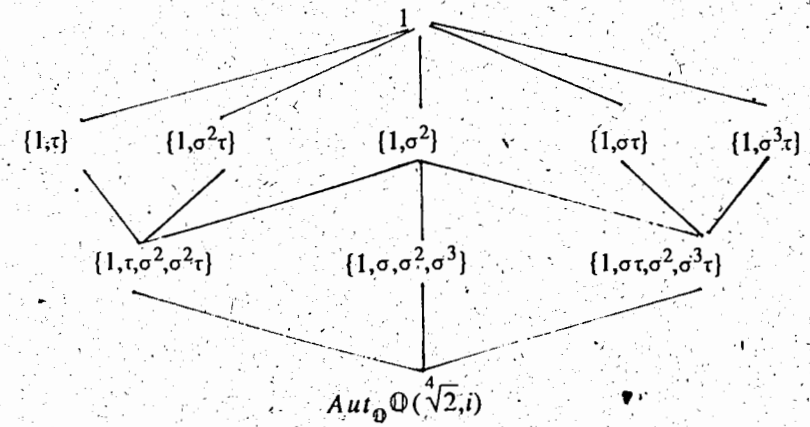
$$\varphi_5: \sqrt[4]{2} \to -\sqrt[4]{2}, \qquad i \to i,$$

$$\varphi_6: \sqrt[4]{2} \to -\sqrt[4]{2}, \qquad i \to -i,$$

$$\varphi_7: \sqrt[4]{2} \to -\sqrt[4]{2}i, \qquad i \to i,$$

$$\varphi_8: \sqrt[4]{2} \to -\sqrt[4]{2}i, \qquad i \to -i.$$

We put $\varphi_2 = \tau$ and $\varphi_3 = \sigma$. Then $o(\tau) = 2$, $o(\sigma) = 4$ and $\sigma^\tau = \sigma^{-1}$. Thus $G(\mathbb{Q}(\sqrt[4]{2},i)/\mathbb{Q})$ is a dihedral group of order 8. Since any automorphism in $G(\mathbb{Q}(\sqrt[4]{2},i)/\mathbb{Q})$ is completely determined by its effect on the four roots $u_1 = \sqrt[4]{2}$, $u_2 = \sqrt[4]{2}i$, $u_3 = -\sqrt[4]{2}$, $u_4 = -\sqrt[4]{2}i$ of $x^4 - 2$, the group $G(\mathbb{Q}(\sqrt[4]{2},i)/\mathbb{Q})$ is isomorphic to a subgroup of S_4. We see $\sigma = \begin{pmatrix} u_1 & u_2 & u_3 & u_4 \\ u_2 & u_3 & u_4 & u_1 \end{pmatrix} = (u_1u_2u_3u_4)$ and $\tau = \begin{pmatrix} u_1 & u_2 & u_3 & u_4 \\ u_1 & u_4 & u_3 & u_2 \end{pmatrix} = (u_2u_4)$. So $G(\mathbb{Q}(\sqrt[4]{2},i)/\mathbb{Q}) \cong \langle (24),(1234) \rangle = \{\iota,(13),(24),(12)(34),(13)(24),(14)(23),(1234),(1432)\} \leqslant S_4$ by an isomorphism $\varphi_2 = \tau \to (24)$, $\varphi_3 = \sigma \to (1234)$. The subgroups of $G(\mathbb{Q}(\sqrt[4]{2},i)/\mathbb{Q})$ are



Let us find the intermediate field of $\mathbb{Q}(\sqrt[4]{2},i)/\mathbb{Q}$ corresponding to $\{1,\sigma^2\tau\}$. We write $u = \sqrt[4]{2}$ for brevity. We have $(u)\sigma^2\tau = (u\sigma)\sigma\tau = (ui)\sigma\tau = (u\sigma\cdot i\sigma)\tau = (ui\cdot i)\tau = (-u)\tau = -(u\tau) = -u$ and $(i)\sigma^2\tau = (i\sigma)\sigma\tau = (i\sigma)\tau = i\tau = -i$. Now let $a,b,c,d,e,f,g,h \in \mathbb{Q}$ and $s = a + bu + cu^2 + du^3 + ei + fui + gu^2i + hu^3i$. Then

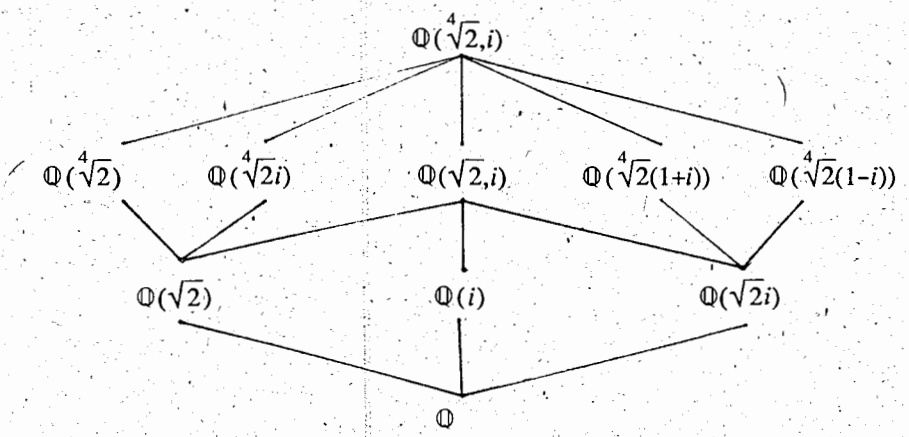
$$\begin{aligned} s\sigma^2\tau &= (a + bu + cu^2 + du^3 + ei + fui + gu^2i + hu^3i)\sigma^2\tau \\ &= a + b(-u) + c(-u)^2 + d(-u)^3 + e(-i) + f(-u)(-i) + g(-u)^2(-i) + h(-u)^3(-i) \\ &= a - bu + cu^2 - du^3 - ei + fui - gu^2i + hu^3i \end{aligned}$$

and so s is fixed under $\sigma^2\tau$ if and only if

$$a = a, \quad b = -b, \quad c = c, \quad d = -d,$$
$$e = -e, \quad f = f, \quad g = -g, \quad h = h,$$

so if and only if $b = d = e = g = 0$,
so if and only if $s = a + cu^2 + fui + hu^3i = a + f(ui) - c(ui)^2 - h(ui)^3$
so if and only if $s \in \mathbb{Q}(ui)$.

Thus the intermediate field of $\mathbb{Q}(\sqrt[4]{2},i)/\mathbb{Q}$ corresponding to $\{1,\sigma^2\tau\}$ is $\{1,\sigma^2\tau\}' = \mathbb{Q}(ui) = \mathbb{Q}(\sqrt[4]{2}i)$. Similar computations yield that the Galois correspondence is as in the diagram below, where intermediate fields occupy the same relative position as the corresponding subgroups.



(c) Let p be a prime number and $n \in \mathbb{N}$. We consider the extension $\mathbb{F}_{p^n}/\mathbb{F}_p$. The mapping

$$\sigma: \mathbb{F}_{p^n} \to \mathbb{F}_{p^n}$$
$$a \to a^p$$

is a field homomorphism (Lemma 52.2) and fixes every element in $\mathbb{F}_p$ (Theorem 12.7 or Theorem 52.8). Thus σ is $\mathbb{F}_p$-linear and, since $|\mathbb{F}_{p^n}:\mathbb{F}_p|$ is finite, σ is onto $\mathbb{F}_{p^n}$ (Theorem 42.22). So σ is an $\mathbb{F}_p$-automorphism of $\mathbb{F}_{p^n}$.

We put $G = Aut_{\mathbb{F}_p}\mathbb{F}_{p^n}$. We want to show $G = \langle\sigma\rangle$. First we prove $o(\sigma) = n$. From $a\sigma^n = a^{p^n} = a$ for all $a \in \mathbb{F}_{p^n}$ (Lemma 52.4(2) or Theorem 52.8), we get $\sigma^n = 1$, so $o(\sigma)|n$. On the other hand, if m is a positive proper divisor of n, then $\mathbb{F}_{p^n}$ has a proper subfield $\mathbb{F}_{p^m}$ with p^m elements (Theorem 52.8) and there is a $b \in \mathbb{F}_{p^n}\backslash\mathbb{F}_{p^m}$ with $b\sigma^m = b^{p^m} \neq b$, so $\sigma^m \neq 1$. So we conclude $o(\sigma) = n$. Since $|\mathbb{F}_{p^n}:\mathbb{F}_p|$ is finite, we get

$$n = o(\sigma) = |\langle\sigma\rangle| \leqslant |G| = |G:1| = |(\mathbb{F}_p)':(\mathbb{F}_{p^n})'| \leqslant |\mathbb{F}_{p^n}:\mathbb{F}_p| = n$$

from Lemma 54.14, so $|\langle\sigma\rangle| = |G| = n$ and $G = \langle\sigma\rangle$.

It is now easy to show that $\mathbb{F}_{p^n}$ is Galois over $\mathbb{F}_p$. We have

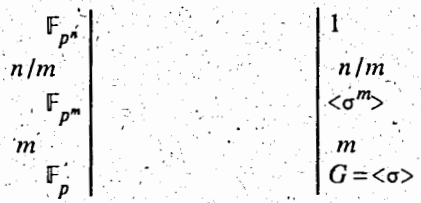
$$G' = \langle\sigma\rangle' = \{a \in \mathbb{F}_{p^n} : a\sigma = a\} = \mathbb{F}_p$$

by Theorem 52.8 and thus $\mathbb{F}_{p^n}$ is Galois over $\mathbb{F}_p$.

The Galois correspondence is easy to describe. The subgroups of $\langle\sigma\rangle$ are in one-to-one correspondence with the positive divisors of n and any subgroup H of G is of the form $H = \langle\sigma^m\rangle$ (Theorem 11.8). The subfield of $\mathbb{F}_{p^n}$ corresponding to $H = \langle\sigma^m\rangle$ is

$$H' = \langle\sigma^m\rangle' = \{a \in \mathbb{F}_{p^n} : a\sigma^m = a\} = \{a \in \mathbb{F}_{p^n} : a^{p^m} = a\} = \mathbb{F}_{p^m},$$

the unique subfield of $\mathbb{F}_{p^n}$ with p^m elements.



In all these examples, we first determined the subgroups of the Galois group and then found the intermediate fields corresponding to them. One can of course reverse this, i.e., one can determine the intermediate fields in the first place and then find the subgroups corresponding to them. However, it is in general more difficult to find all intermediate fields of an extension, for it is likely that one overlooks some of them. Also, it is more difficult to avoid duplications. For instance, in Example

54.18(c), it is not immediately clear where $\mathbb{Q}(\sqrt{2}(1+i))$ and $\mathbb{Q}(\sqrt{2}(1-i))$ are, nor whether $\mathbb{Q}(\sqrt{2}(1+i)) = \mathbb{Q}(\sqrt{2}(1-i))$. It is far easier to list the subgroups than to list the intermediate fields.

It is natural to ask which intermediate fields correspond to the normal subgroups of the Galois group of an extension. Also, what can be said about the factor groups of the Galois group? We proceed to answer these questions. We need a definition.

54.19 Definition: Let E/K be a field extension and let $G = Aut_K E$ be its Galois group. An intermediate field L of this extension is said to be *stable relative to K and E*, or to be *(K,E)-stable* if every K-automorphism $\varphi \in Aut_K E$ of E maps L into L.

In the situation of Definition 54.19, if L is a (K,E)-stable intermediate field, then the inverse φ^{-1} of any K-automorphism φ of E also maps L into L. Thus the restriction $\varphi_{|L}$ to L of any K-automorphism of E is a K-automorphism of L. Thus we have a "restriction" mapping

$$\begin{aligned} res: Aut_K E &\to Aut_K L \\ \varphi &\to \varphi_{|L} \end{aligned}$$

A K-automorphism λ of L is said to be *extendible to E* if there is a K-automorphism φ of E such that $\lambda = \varphi_{|L}$. Therefore *res* is a mapping onto the set of all extendible K-automorphisms of L.

54.20 Theorem: *Let E/K be a field extension.*
(1) *If L is a (K,E)-stable intermediate field, then L' is a normal subgroup of the Galois group $Aut_K E$.*
(2) *If H is a normal subgroup of $Aut_K E$, then H' is a (K,E)-stable intermediate field of the extension.*

Proof: (1) We are to prove that $\varphi^{-1}\lambda\varphi \in L'$ for all $\lambda \in L'$ and $\varphi \in Aut_K E$. Thus we must show that $a(\varphi^{-1}\lambda\varphi) = a$ for all $a \in L$. Indeed, if $a \in L$, $\lambda \in L'$ and $\varphi \in Aut_K E$, then $a\varphi^{-1} \in L$ since L is (K,E)-stable, so $(a\varphi^{-1})\lambda = a\varphi^{-1}$, so $a(\varphi^{-1}\lambda\varphi) = (a\varphi^{-1}\lambda)\varphi = (a\varphi^{-1})\varphi = a$. Hence $L' \trianglelefteq Aut_K E$.

(2) We are to prove that $a\varphi \in H'$ for all $a \in H'$ and $\varphi \in Aut_K E$. Thus we must show that $(a\varphi)\eta = a\varphi$ for all $\eta \in H$. Indeed, if $a \in H'$, $\eta \in H$ and $\varphi \in Aut_K E$, then $\varphi\eta\varphi^{-1} \in H$ since $H \trianglelefteq Aut_K E$, so $a(\varphi\eta\varphi^{-1}) = a$, so $a(\varphi\eta)\varphi^{-1} = a$, so $a(\varphi\eta) = a\varphi$. Hence H' is (K,E)-stable. □

54.21 Theorem: *Let E/K be a Galois extension and L an intermediate field. If L is (K,E)-stable, then L is Galois over K.*

Proof: For any $a \in L\backslash K$, we must find a $\lambda \in Aut_K L$ such that $a\lambda \neq a$. Since E is Galois over K, there is a $\varphi \in Aut_K E$ such that $a\varphi \neq a$. Then $\varphi_{|L} \in Aut_K L$ by stability of L relative to K and E. Thus $\varphi_{|L}$ can be taken as λ. □

54.22 Theorem: *Let E/K be a Galois extension and $f(x) \in K[x]$ be irreducible in $K[x]$. If $f(x)$ has a root in E, then $f(x)$ splits in E and the roots of $f(x)$ are all simple.*

Proof: Let a_1 be a root of $f(x)$ in E. We put $deg\ f(x) = n$. We want to show that $f(x) = c(x - a_1)(x - a_2)\ldots(x - a_n)$ for some elements $c,a_1,a_2,\ldots,a_n$ in E. For this purpose, we put $g(x) = (x - a_1)(x - a_2)\ldots(x - a_m) \in E[x]$, where $a_1,a_2,\ldots,a_m$ are all the distinct roots of $f(x)$ in E. We know $m \leqslant n$ from Theorem 35.7.

Any K-automorphism of E maps a root of $f(x)$ to a root of $f(x)$ (Lemma 54.5). Thus the coefficients of $g(x)$, which are symmetric in the roots $a_1,a_2,\ldots,a_m$ of $g(x)$, are fixed by any K-automorphism of E. This shows that the coefficients of $g(x)$ are in $E' = K$. Hence $g(x) \in K[x]$. Then $f(x)$ and $g(x)$ are two polynomials in $K[x]$ with a common root a_1 and $f(x)$ is irreducible over K. Theorem 35.18(1),(3) gives then $f(x)|g(x)$ and consequently $n = deg\ f(x) \leqslant deg\ g(x) = m$. We have $m \leqslant n$ also, thus $n = m$. From $f(x)|g(x)$ we get then $f(x) \approx g(x)$. So $f(x) = c(x - a_1)(x - a_2)\ldots(x - a_n)$ for some $c \in K^*$ and the roots $a_1,a_2,\ldots,a_n \in E$ of are all distinct, i.e., all roots of $f(x)$ are simple. □

The next theorem is a kind of converse to Theorem 54.21. The result is not necessarily true without the hypothesis that L is algebraic (cf. Ex. 8).

54.23 Theorem: *Let E/K be a field extension and L an intermediate field. If L is algebraic and Galois over K, then L is (K,E)-stable.*

Proof: We want to show that $a\varphi \in L$ for any $a \in L$ and any $\varphi \in Aut_K E$. If $a \in L$, then a is algebraic over K since L is algebraic over K. Let $f(x)$ be the minimal polynomial of a over K. Then $f(x)$ is a product of n distinct polynomials of degree one in $L[x]$ because L is Galois over K (Theorem 54.22). Thus all roots of $f(x)$ are in L. Now if $\varphi \in Aut_K E$, then $a\varphi$ is a root of $f(x)$, hence $a\varphi \in L$, as was to be proved. □

Let E/K be a field extension and let L be a (K,E)-stable intermediate field of E/K. Let us consider the restriction mapping

$$res: Aut_K E \to Aut_K L$$
$$\varphi \to \varphi_{|L}$$

Since $(\varphi\psi)_L = \varphi_L \psi_L$ for any two K-automorphisms φ,ψ of E, we see that *res* is a homomorphism. Therefore $(Aut_K E)/Ker\ res = Im\ res$. Now $Im\ res$ is the set of all K-automorphisms of L that are extendible to E (hence the set of all K-automorphisms of L that are extendible to E is a subgroup of $Aut_K E$) and $Ker\ res = \{\varphi \in Aut_K E: \varphi_{|L} = i_L\} = \{\varphi \in Aut_K E: a\varphi = a \text{ for all } a \in L\} = L' = Aut_L E$. Hence $(Aut_K E)/(Aut_L E)$ is isomorphic to the group of all K-automorphisms of L that are extendible to E. We proved the

54.24 Theorem: *Let E/K be a field extension and L an intermediate field. If L is (K,E)-stable, then ($L' = Aut_L E$ is normal in $Aut_K E$ and) the quotient group $G(E/K)/G(E/L) = (Aut_K E)/(Aut_L E)$ is isomorphic to the subgroup of $Aut_K L$ consisting exactly of the K-automorphisms of L that are extendible to E.* □

We can now supplement the fundamental theorem by describing the situation with respect to an intermediate field.

54.25 Theorem: *Let E/K be a finite dimensional Galois extension of fields and $G = Aut_K E$. Let L be an intermediate field of E/K.*

(1) *E is Galois over L.*
(2) *L is Galois over K if and only if* $L' = Aut_L E$ *is normal in* $G = Aut_K E$. *In this case,* $G/L' = (Aut_K E)/(Aut_L E)$ *is isomorphic to the Galois group* $Aut_K L$ *of L over K. Thus* $G(E/K)/G(E/L) \cong G(L/K)$.

Proof: Here the hypotheses of the fundamental theorem are satisfied. The fundamental theorem states that any intermediate field of E/K and any subgroup of G is closed.

(1) In order to show that E is Galois over L, we must prove that $L = L''$, that is, that L is closed. This follows from the fundamental theorem.

(2) E/K is a finite dimensional extension by hypothesis and so L/K is also a finite dimensional extension. Thus L is algebraic over K (Theorem 50.10) If L is Galois over K, then L is (K,E)-stable by Theorem 54.24 and so L' is normal in $Aut_K E$ by Theorem 54.20(1). Conversely, if L' is normal in $Aut_K E$, then L'' is a (K,E)-stable intermediate field by Theorem 54.20(2). Here $L = L''$ because all intermediate fields are closed. Thus L is (K,E)-stable. Theorem 54.21 tells then that L is Galois over K. So L is Galois over K if and only if L' is normal in $G = Aut_K E$.

Suppose now L is Galois over K and $L' \trianglelefteq G = Aut_K E$. Then $|Aut_K L| = |L:K|$ by the fundamental theorem (with L in place of E). Theorem 54.23 states that $G/L' = (Aut_K E)/(Aut_L E)$ is isomorphic to a subgroup of $Aut_K L$. Using $L = L''$ (i.e., L is closed) and $G' = K$ (i.e., L is Galois over K), we see $|G/L'| = |G:L'| = |L'':G'| = |L:K| = |Aut_K L|$ by the fundamental theorem. Thus G/L', which is isomorphic to a subgroup of $Aut_K L$, has the same order as $Aut_K L$. Since $|Aut_K L| = |L:K|$ is finite, this implies that G/L' is actually isomorphic to $Aut_K L$ itself, as was to be shown. □

We end this paragraph with an important illustration of Theorem 54.25.

54.26 Theorem: *Let* $\mathbb{F}_q$ *be a field of q elements and E a finite dimensional extension of* $\mathbb{F}_q$. *Then E is Galois over* $\mathbb{F}_q$ *and* $Aut_{\mathbb{F}_q} E$ *is cyclic, generated by the automorphism* φ, *where* $\varphi: a \to a^q$ *for all* $a \in E$.

Proof: Let $|E:\mathbb{F}_q| = r$ and $char\ \mathbb{F}_q = p$, so that $\mathbb{F}_p$ is the prime subfield of $\mathbb{F}_q$ (and of E). We have $q = p^m$, where $m = |\mathbb{F}_q:\mathbb{F}_p|$. We consider the extension $E/\mathbb{F}_p$. Since E is an r-dimensional vector space over $\mathbb{F}_q$ and $\mathbb{F}_q$ is an m-dimensional vector space over $\mathbb{F}_p$, Theorem 48.13 says E is an rm-dimensional vector space over $\mathbb{F}_p$ and so $|E| = p^{rm}$. Thus E is a finite field and E is Galois over $\mathbb{F}_p$ (Example 54.18(c)). Then E is Galois over any intermediate field of $E/\mathbb{F}_p$ (Theorem 54.25(1)); in particular, E is Galois over $\mathbb{F}_q$. Furthermore, we know from Example 54.18(c) that $Aut_{\mathbb{F}_p}E = \langle\sigma\rangle$, where σ is the field isomorphism $a \to a^p$ for all $a \in E$ and that the group $(\mathbb{F}_q)'$ corresponding to the intermediate field $\mathbb{F}_q$ with p^m elements is $\langle\sigma^m\rangle$. Thus $Aut_{\mathbb{F}_q}E = (\mathbb{F}_q)' = \langle\varphi\rangle$, where $\varphi = \sigma^m$ is the mapping $a \to a^{p^m} = a^q$ for all $a \in E$. □

Exercises

1. Find the Galois group $Aut_K E$ and all its subgroups and describe the Galois correspondence between the subgroups of $Aut_K E$ and the intermediate fields of E/K when

   (a) $E = \mathbb{Q}(\sqrt{2},\sqrt{3})$ and $K = \mathbb{Q}$;

   (b) $E = \mathbb{Q}(\sqrt[3]{2},\sqrt[3]{5})$ and $K = \mathbb{Q}$, $K = \mathbb{Q}(\sqrt[3]{2})$;

   (c) $E = \mathbb{Q}(\sqrt[3]{2},\sqrt[4]{3},i)$ and $K = \mathbb{Q}(i)$, $K = \mathbb{Q}(i,\sqrt[4]{3})$;

   (d) $E = \mathbb{Q}(\sqrt[3]{2},i)$ and $K = \mathbb{Q}(i)$;

   (e) $E = \mathbb{Q}(\sqrt{2},\sqrt[3]{5})$ and $K = \mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$.

2. Let E/K be a field extension. Prove that if L is a (K,E)-stable intermediate field, so is L'' and that if H is a normal subgroup of $Aut_K E$, so is H''.

3. Let E/K be a field extension and $G = Aut_K E$. Let L, M be intermediate fields of E/K and let H, J be subgroups of G. Prove that $\langle H \cup J\rangle' = H' \cap J'$ and $(LM)' = L' \cap M'$.

If, in addition, L is finite dimensional and Galois over K, then LM is finite dimensional and Galois over M and $Aut_{L\cap M}L \cong Aut_M LM$.

4. Let K be a field and x an indeterminate over K. Show that, if L is an intermediate field of $K(x)/K$ and $L \neq K$, then $|K(x):L|$ is finite.

5. Prove that $K(x)$ is Galois over K if and only if K is infinite.

6. Let K be an infinite field. Prove that a proper subgroup of $Aut_K K(x)$ is closed if and only if it is a finite subgroup of $Aut_K K(x)$.

7. Consider the extension $\mathbb{Q}(x)/\mathbb{Q}$. Prove that the intermediate field $\mathbb{Q}(x^2)$ is closed and the intermediate field $\mathbb{Q}(x^3)$ is not closed.

8. Let K be an infinite field and x,y two distinct indeterminates over K. Show that the intermediate field $K(x)$ of the extension $K(x,y)/K$ is Galois over K but $K(x)$ is not stable relative to K and $K(x,y)$.

§55
Separable Extensions

In §54, we established the foundations of Galois theory, but we have no handy criterion for determining whether a given field extension is Galois or not. Even in the quite simple cases such as in Example 54.18, we had to study the effects of automorphisms on the elements in the extension field, and this involved much calculation. The extension fields in Example 54.18 were seen to be splitting fields of certain polynomials over the base field. In this paragraph, we will learn that a finite dimensional extension is Galois if and only if the extension field is a splitting field of a polynomial whose irreducible factors have no multiple roots. We give a name to irreducible polynomials of this kind.

55.1 Definition: Let K be a field and $f(x) \in K[x]$. If $f(x)$ is irreducible over K and has no multiple roots (in any splitting field of $f(x)$ over K), then $f(x)$ is said to be *separable over* K.

Thus all the *deg* $f(x)$ roots of a polynomial $f(x)$ separable over K are distinct and $f(x)$ splits into distinct linear factors in any splitting field of $f(x)$ over K.

The existence of multiple roots can be decided by means of the derivative. If K is a field, $f(x)$ an irreducible polynomial in $K[x]$ and E a splitting field of $f(x)$ over K, then Theorem 35.18 (5) and Theorem 35.18 (6) show that $f(x)$ is separable over K if and only if $f'(x) \neq 0$.

How can an irreducible polynomial $f(x)$ have a zero derivative? Now $f(x)$ is not 0 or a unit because of irreducibility, so *deg* $f(x) =: m \geqslant 1$. Let $f(x) = \sum_{i=0}^{m} a_i x^i$, with $a_m \neq 0$. Then $f(x) = \sum_{i=1}^{m} i\, a_i x^{i-1} = 0$ if and only if $i a_i = 0$ for all $i = 1,2, \ldots, m$. In particular, $(m1)a_m = m\, a_m = 0$. Since a field has no zero divisors and $a_m \neq 0$, this forces $m1 = 0$. This is impossible in case *char* $K = 0$ and is equivalent to $p|m$ in case *char* $K = p \neq 0$. Likewise, if $a_i \neq 0$, the

condition $ia_i = 0$ is equivalent to $p|i$ in case *char* $K = p$. So for terms $a_i x^i$ with $a_i \neq 0$, we have $i = pj$ for some j and we may write $f(x) = \sum_{j=0}^{[m/p]} a_{pj} x^{pj}$.

Putting $[m/p] = n$, $a_{pj} = b_j$ and $g(x) = \sum_{j=0}^{n} b_j x^j$, we obtain $f(x) = g(x^p)$. Thus $f(x)$ is actually a polynomial in x^p. Conversely, if $f(x) = g(x^p)$, then $f'(x) = g'(x^p) \cdot px^{p-1} = g'(x^p) \cdot 0 = 0$ by Lemma 35.16. We summarize:

55.2 Lemma: *Let K be a field. If char $K = 0$, then any polynomial irreducible over K is separable over K. If char $K = p \neq 0$ and $f(x) \in K[x]$ is irreducible over K, then $f(x)$ is separable over K if and only if $f(x)$ is not a polynomial in x^p, i.e., $f(x)$ is not separable over K if and only if $f(x) = g(x^p)$ for some $g(x) \in K[x]$.* □

In terms of separable polynomials we now define separable *elements* and separable field *extensions*.

55.3 Definition: Let E/K be a field extension and $a \in E$. If a is algebraic over K and the minimal polynomial of a over K is separable over K, then a is said to be *separable over* K.

Thus any element a of K is separable over K since the minimal polynomial of a over K is $x - a \in K[x]$ and $x - a$ is separable over K.

55.4 Definition: Let E/K be a field extension. If E is algebraic over K and if every element of E is separable over K, then E is said to be *separable over* K or a *separable extension of* K and E/K is called a *separable extension*.

The polynomial $x^2 + 1 \in \mathbb{Q}[x]$ is separable over $\mathbb{Q}$, because it is irreducible over $\mathbb{Q}$ and *char* $\mathbb{Q} = 0$. On the other hand, $x^2 + 1 \in \mathbb{F}_2[x]$ is not separable over $\mathbb{F}_2$ because $x^2 + 1 = (x + 1)^2$ is not even irreducible over $\mathbb{F}_2$.

If E/K is an extension of fields of characteristic 0, then any element of E that is algebraic over K is separable over K. Thus any algebraic extension of a field of characteristic 0 is a separable extension of that field.

We compare separability over a field with separability over an intermediate field.

54.5 Lemma: *Let E/K be a field extension and let L be an intermediate field of E/K. Let $a \in E$ be algebraic over K. If a is separable over K, then a is separable over L.*

Proof: Lemma 50.5 shows that a is algebraic over L. Let $f(x)$ be the minimal polynomial of a over K and $g(x)$ the minimal polynomial of a over L. By Lemma 50.5, $g(x)$ is a divisor of $f(x)$. Thus any root of $g(x)$ is a root of $f(x)$. Since a is separable over K, the roots of $f(x)$ are all simple, hence, all the more so, the roots of $g(x)$ are all simple and a is separable over L. □

55.6 Lemma: *Let E/K be a field extension and let L be an intermediate field of E/K. Then E is separable over L and L is separable over K.*

Proof: Assume that E is separable over K. We are to show that (1) E is algebraic over L and L is algebraic over K and (2) any element of E is separable over L and any element of L is separable over K. Since E is separable over K, we deduce E is algebraic over L (Lemma 50.5) and any element of E, being (algebraic and) separable over K, is also separable over L (Lemma 55.5). Thus E/L is a separable extension. Moreover, all elements of E are separable over K, so, in particular, all elements in L are separable over K and L/K is a separable extension. □

The converse of Lemma 55.6 is also true and will be proved later in this paragraph (Theorem 55.19). Our next goal is to characterize Galois extensions as splitting fields of separable polynomials.

55.7 Theorem: *Let E/K be a finite dimensional field extension. Then the following statements are equivalent.*

(1) *E is Galois over K.*

(2) *E is a separable extension of K and the splitting field over K of a polynomial in $K[x]$.*

(3) *E is the splitting field of a polynomial in $K[x]$ whose irreducible factors are separable over K.*

Proof: (1) $\Rightarrow$ (2) We prove E/K is a separable extension. Since E/K is a finite dimensional extension, E is algebraic over K. We have also to show that the minimal polynomial over K of any element u in E is separable over K. This follows immediately from Theorem 54.22. Hence E is a separable extension of K.

We must now show that there is a polynomial $g(x)$ in $K[x]$ such that E is a splitting field of $f(x)$ over K. Let $\{a_1,a_2, \ldots,a_m\}$ be a K-basis of E and let $f_i(x) \in K[x]$ be the minimal polynomial of a_i over K $(i = 1,2, \ldots,m)$. We put $g(x) = f_1(x)f_2(x)\ldots f_m(x) \in K[x]$. From Theorem 54.22 again, we learn that each $f_i(x)$, hence also $g(x)$, splits in E. Moreover, $g(x)$ cannot split in any *proper* subfield L of E containing K for if L is an intermediate field of E/K and $g(x)$ splits in L, then L contains all roots of $g(x)$, hence L contains $a_1,a_2, \ldots,a_m$ and we have

$$E = s_K(a_1,a_2, \ldots,a_m) \subseteq K(a_1,a_2, \ldots,a_m) \subseteq L,$$

so $E = L$. Thus E is indeed a splitting field of $g(x)$ over K.

(2) $\Rightarrow$ (3) Assume now E is separable over K and E is a splitting field over K of a polynomial $g(x)$ in $K[x]$. We are to prove that the irreducible factors of $g(x)$ in $K[x]$ are separable over K. Let $g(x) = f_1(x)f_2(x)\ldots f_m(x)$ be the decomposition of $g(x)$ into irreducible factors $f_i(x)$ in $K[x]$. Since $g(x)$ splits in E, each $f_i(x)$ has a root $a_i \in E$. Here a_i is separable over K because E is separable over K. Thus the minimal polynomial of a_i over K is a separable polynomial over K. But the minimal polynomial of a_i over K is $c_if_i(x)$ with some suitable $c_i \in K$, because a_i is a root of $f_i(x)$ and $f_i(x)$ is

irreducible in $K[x]$. So $c_i f_i(x)$ is separable over K and consequently $f_i(x)$ is also separable over K.

(3) $\Rightarrow$ (1) Suppose now E is a splitting field of a polynomial $g(x) \in K[x]$ whose irreducible factors in $K[x]$ are separable over K. We put

$$K_0 = \{a \in E: a\varphi = a \text{ for all } \varphi \in Aut_K E\}.$$

Clearly $K_0 \subseteq K$. In fact K_0 is the fixed field of $Aut_K E$, hence K_0 is an intermediate field of the extension E/K. We prove that E is Galois over K by showing (i) E is Galois over K_0; (ii) $Aut_K E = Aut_{K_0} E$; (iii) $|E:K| = |Aut_K E|$. These will indeed imply

$$|E:K_0| = |Aut_{K_0} E| \qquad \text{(by the fundamental theorem of Galois theory, since } E/K_0 \text{ is a finite dimensional Galois extension),}$$

$$|Aut_{K_0} E| = |Aut_K E| \qquad \text{(by (ii)),}$$

$$|Aut_K E| = |E:K| \qquad \text{(by (iii)),}$$

so $|E:K_0| = |E:K|$,

so $K_0 = K$

and E is Galois over K (by (i)).

Since, for any $\varphi \in Aut_K E$, there holds $a\varphi = a$ for all $a \in K_0$, we see that $Aut_K E \leqslant Aut_{K_0} E$.

(i) In order to show that E is Galois over K_0, we have to find, for each $b \in E \backslash K_0$, an automorphism $\varphi \in Aut_{K_0} E$ such that $b\varphi \neq b$. If $b \in E \backslash K_0$, then, by definition of K_0, there is a $\varphi \in Aut_K E$ such that $b\varphi \neq b$. From $Aut_K E \leqslant Aut_{K_0} E$, we see $\varphi \in Aut_{K_0} E$ and $b\varphi \neq b$. Thus E is Galois over K_0.

(ii) E/K is a finite dimensional extension, hence E/K_0 is a finite dimensional extension and E/K_0 is Galois. Therefore, by the fundamental theorem of Galois theory, the subgroup $Aut_K E$ of $Aut_{K_0} E$ is a *closed* subgroup of $Aut_{K_0} E$. Hence $Aut_{K_0} E = K_0' = ((Aut_K E)')' = (Aut_K E)'' = Aut_K E$.

(iii) We prove $|E:K| = |Aut_K E|$ by induction on $n = |E:K|$, the hypothesis being that E be a splitting field over K of a polynomial in $K[x]$ whose irreducible factors (in $K[x]$) are separable over K.

If $n = 1$, then $E = K$, so $Aut_K E = Aut_K K = \{\iota_K\}$ and $|E:K| = 1 = |\{\iota_k\}| = |Aut_K E|$.

Suppose now $n \geqslant 2$ and suppose that $|E_1:K_1| = |Aut_{K_1} E_1|$ whenever E_1/K_1 is a finite dimensional extension with $1 \leqslant |E_1:K_1| < n$ such that E_1 is a splitting field of a polynomial in $K_1[x]$ whose irreducible factors (in $K_1[x]$) are separable over K_1.

Let $g(x) \in K[x]$ be the polynomial of which E is a splitting field over K and let $g(x) = f_1(x)f_2(x)\ldots f_m(x)$ be the decomposition of $g(x)$ into irreducible polynomials $f_i(x)$ in $K[x]$. The polynomials $f_i(x)$ cannot all be of first degree, for then the roots of $f_i(x)$ would be in K and, as E is a splitting field of $g(x)$ over K, the field E would coincide with K, against the hypothesis $|E:K| = n > 1$. Thus at least one of $f_i(x)$ have degree > 1. Let us assume $deg\ f_1(x) = r > 1$ and let $a \in E$ be a root of $f_1(x)$. We put $L = K(a)$. Then $|L:K| = r$ and $|E:L| = n/r < n$.

Now E is a splitting field of $g(x) \in L[x]$ over L (Example 53.5(e)) and the irreducible factors (in $L[x]$) of $g(x)$, being divisors of $f_i(x)$, have no multiple roots and are therefore separable over L. Since $|E:L| = n/r < n$, we get $|E:L| = |Aut_L E| = |L'|$ by induction.

In order to prove $|E:K| = |Aut_K E|$, i.e., in order to prove $|E:L||L:K| = |Aut_K E:L'||L'|$, it will be thus sufficient to show that $r = |L:K| = |Aut_K E:L'|$.

We show $|Aut_K E:L'| = r$ by defining a one-to-one mapping A from the set $\mathcal{R}$ of right cosets of L' in $Aut_K E$ onto the set of distinct roots of $f_1(x)$ in E. Let $\{a = a_1, a_2, \ldots, a_r\}$ be the distinct roots of $f_1(x)$ in E. We put

$$\begin{aligned} A:\ \mathcal{R} &\to \{a_1, a_2, \ldots, a_r\} \\ L'\varphi &\to a\varphi \end{aligned}$$

($\varphi \in Aut_K E$; we know $a\varphi \in E$ is a root of $f_1(x)$ from Lemma 54.5). This mapping A is well defined, for if $L'\varphi = L'\psi$, then

$$\begin{aligned} &\varphi\psi^{-1} \in L' \\ &\varphi\psi^{-1} \text{ fixes each element of } L = K(a) \\ &\varphi\psi^{-1} \text{ fixes } a \\ &a(\varphi\psi^{-1}) = a \\ &(a\varphi)\psi^{-1} = a \\ &a\varphi = a\psi \\ &(L'\varphi)A = (L'\psi)A, \end{aligned}$$

so A is well defined and, reading the lines backwards, we see that A is one-to-one as well. It remains to show that A is onto. Indeed, if a_i is any root of $f_1(x)$ in E, then there is a field homomorphism $\alpha_i: K(a) \to K(a_i)$

mapping a to a_i and fixing each element of K (Theorem 53.1) and α_i can be extended to a K-automorphism $\varphi_i: E \to E$ (Theorem 53.7). Then A sends the coset $L\varphi_i \in \mathcal{R}$ to $a\varphi_i = a\alpha_i = a_i$. Hence A is onto. This gives $|Aut_K E:L| = |\mathcal{R}| = |\{a_1, a_2, \ldots, a_r\}| = r$. The proof is complete. □

Thus for finite dimensional extensions, being Galois is equivalent to separability plus being a splitting field.

If E/K is a field extension and E is a splitting field of $f(x) \in K[x]$ over K, then all roots of the polynomial $f(x)$ are in E. We show more generally that, if there is a root in E of a polynomial over K, then all roots of that polynomial are in E. This gives a characterization of splitting fields without referring to any particular polynomial.

55.8 Theorem: *Let E/K be a finite dimensional field extension. The following statements are equivalent.*
(1) *There is a polynomial $f(x) \in K[x]$ such that E is a splitting field of $f(x)$ over K.*
(2) *If $g(x)$ is any irreducible polynomial in $K[x]$, and if $g(x)$ has a root in E, then $g(x)$ splits in E.*

Proof: (1) $\Longrightarrow$ (2) Asume that $g(x) \in K[x]$ is irreducible over K and that $g(x)$ has a root $u \in E$. We want to show that all irreducible factors of $g(x)$ in $E[x]$ have degree one. Suppose, on the contrary, that $h(x) \in E[x]$ is an irreducible (over E) factor of $g(x)$ with *deg* $h(x) = n > 1$. We adjoin a root t of $h(x)$ to E and thereby construct the field $E(t)$.

Now u and t are roots of the irreducible polynomial $g(x)$ in $K[x]$, so there is a K-isomorphism $\varphi: K(u) \to K(t)$ (Theorem 53.2). Since E is a splitting field of $f(x)$ over $K(u)$ and $E(t)$ is a splitting field of $f(x)$ over $K(t)$ (Example 53.5(e)), the K-isomorphism φ can be extended to a K-isomorphism $\psi: E \to E(t)$ (Theorem 53.7). But then $|E:K| = |E(t):K| = |E(t):E||E:K| = n|E:K| > |E:K|$, a contradiction. Thus all irreducible factors of $g(x)$ in $E[x]$ have degree one and $g(x)$ splits in E.

(2) $\Longrightarrow$ (1) Suppose now that any irreducible polynomial in $K[x]$ splits in E whenever it has a root in E. Let $\{a_1, a_2, \ldots, a_m\}$ be a K-basis of E and let

$f_i(x) \in K[x]$ be the minimal polynomial of a_i over K. We put $f(x) = f_1(x)f_2(x)\ldots f_m(x)$. We claim E is a splitting field of $f(x)$ over K.

Each $f_i(x)$ has a root a_i in E, so each $f_i(x)$ splits in E by hypothesis, so $f(x)$ splits in E. Moreover, $f(x)$ cannot split in a proper subfield of E containing K for if L is an intermediate field of E/K and $f(x)$ splits in L, then all roots of $f(x)$ will be in L, in particular each a_i will be in L, thus $E = s_K(a_1,a_2,\ldots,a_m) \subseteq K(a_1,a_2,\ldots,a_m) \subseteq L$. Hence E is a splitting field of $f(x)$ over K. □

Theorem 55.8 leads us to

55.9 Definition: Let E/K be a field extension. If E is algebraic over K and if every irreducible polynomial in $K[x]$ that has a root in E in fact splits in E, then E is said to be *normal over* K, and E/K is called a *normal extension*.

With this terminology, Theorem 55.8 reads as follows.

55.8 Theorem: *A finite dimensional extension E/K is a normal extension if and only if E is a splitting field over K of a polynomial in $K[x]$.* □

55.10 Theorem: *Let E/K be a finite dimensional field extension. E is Galois over K if and only if E is normal and separable over K.*

Proof: This is immediate from Theorem 55.7(2) and Theorem 55.8. □

55.11 Theorem: *Let E/K be a finite dimensional field extension. There is an extension field N of E such that*

(i) *N is normal over K;*

(ii) *no proper subfield of N containing E is normal over K;*

(iii) *$|N:K|$ is finite.*

(iv) *N is Galois over K if and only if E is separable over K.*

Moreover, if N' is another extension field of E with the same properties, then N and N' are E-isomorphic.

Proof: Let $\{a_1,a_2, \ldots ,a_m\}$ be a K-basis of E and let $f_i(x) \in K[x]$ be the minimal polynomial of a_i over K. We put $f(x) = f_1(x)f_2(x)\ldots f_m(x) \in K[x]$. Let N be a splitting field of $f(x)$ over E, with $|N:E|$ finite (Theorem 53.6). We claim N has the properties stated above. Since $|N:E|$ and $|E:K|$ are both finite, $|N:K|$ is finite. This proves (iii).

To establish (i), we show that N is a splitting field of $f(x)$ over K (Theorem 55.8). Certainly $f(x)$ splits in N, because N is a splitting field of $f(x)$ over E. Now we have to prove that $f(x)$ does not split in any proper subfield of N containing K. If L is an intermediate field of N/K in which $f(x)$ splits, then L contains all roots of $f(x)$, hence $\{a_1,a_2, \ldots ,a_m\} \subseteq L$, hence $E = s_K(a_1,a_2, \ldots ,a_m) \subseteq K(a_1,a_2, \ldots ,a_m)L \subseteq N$; so L, in which $f(x)$ splits, is an intermediate field of N/E; so $L = E$ since N is a splitting field of $f(x)$ over E. Thus N is indeed a splitting field of $f(x)$ over K.

Now (ii). If L is a proper subfield of N containing E, then L cannot be normal over K. Otherwise L, containing a root a_i of $f_i(x)$, would in fact contain all roots of $f_i(x)$ by normality, hence L would contain all the roots of $f(x)$; thus L would contain E and all roots of $f(x)$. Then L would contain H, where H is the subfield of N generated by the roots of $f(x)$ over E. But H is the unique splitting field of $f(x)$ which is an intermediate field of N/E (Example 53.5(d)), so $N = H \subseteq L$ and this forces $L = N$. This establishes (ii).

(iv) If N is Galois over K, then N is separable over K and the intermediate field E of N/K is also separable over K (Lemma 55.6). Conversely, if E is separable over K, then a_i are separable over K, so $f_i(x)$ are separable over K and N' is a splitting field over K of a polynomial $f(x)$ whose irreducible divisors are separable over K. Thus N is Galois over K (Theorem 55.7).

Finally, let N' be any extension field satisfying (i),(ii),(iii). As $a_i \in E \subseteq N'$ and N' is normal over K, the field N' contains all roots of the minimal polynomial $f_i(x)$ over K, hence N' contains all roots of $f(x)$, hence N' contains a splitting field H' of $f(x)$ over K. Then H' is normal over K (Theorem 55.8). Because of the condition (ii), we get $H' = N'$. Hence N' is a splitting field of $f(x)$ over K. From Example 53.5(e), we deduce that N' is also a splitting

field of $f(x)$ over E. Thus both N and N' are splitting fields of $f(x)$ over E and therefore N and N' are E-isomorphic (Theorem 53.8). □

55.12 Definition: Let E/K be a finite dimensional field extension. An extension field N of E as in Theorem 55.11 is called a *normal closure of E over K.*

Since a normal closure of E over K is unique to within an E-isomorphism, we sometimes speak of *the* normal closure of E over K.

The field $\mathbb{Q}(\sqrt[3]{2},\omega)$ is a normal closure of $\mathbb{Q}(\sqrt[3]{2})$ over $\mathbb{Q}$. Likewise $\mathbb{Q}(\sqrt[4]{2},i)$ is a normal closure of $\mathbb{Q}(\sqrt[4]{2})$ over $\mathbb{Q}$.

Our next topic is the so-called primitive element theorem which states that a finitely generated separable extension is in fact a simple extension. This theorem is due to Abel, but the first complete proof was given by Galois. The elements of a finitely generated separable extension can therefore be expessed in the extremely convenient form $\sum_i a_i u^i$, where u is a primitive element of the extension and a_i are in the base field.

55.13 Theorem: *Let E/K be an algebraic separable extension of fields and $a,b \in E$. Then there is an element c in $K(a,b)$ such that $K(a,b) = K(c)$.*

Proof: We distinguish two cases according as K is a finite or an infinite field.

If K is finite, then $K(a,b)$ is finite dimensional over K (Theorem 50.12) and has $|K|^{[K(a,b):K]}$ elements. Hence $K(a,b)$ is finite and its characteristic is $p \neq 0$, thus $\mathbb{F}_p \subseteq K(a,b)$ and $K(a,b) = \mathbb{F}_p(c)$ for some c in $K(a,b)$ (Theorem 52.19(1); c can be chosen as a generator of the cyclic group $K(a,b)^*$). Then $K(a,b) = K(c)$.

Assume now K is infinite. Let N be a normal closure of E over K (Theorem 55.11). Let $f(x) \in K[x]$ be the minimal polynomial of a over K and $g(x) \in K[x]$ the minimal polynomial of b over K. Since $a,b \in E \subseteq N$ and N is normal over K, all roots of $f(x)$ and $g(x)$ lie in N.

Let $a = a_1,a_2,\ldots,a_n \in N$ be the roots of $f(x)$ and $b = b_1,b_2,\ldots,b_m \in N$ be the roots of $g(x)$. Since E is separable over K, a and b are separable over K, so $f(x)$ and $g(x)$ are separable over K, so $a_i \neq a_j$ when $i \neq j$ $(i,j = 1,2,\ldots,n)$ and $b_k \neq b_l$ when $k \neq l$ $(k,l = 1,2,\ldots,m)$.

There are finitely many elements in N of the form

$$\frac{b_k - b_l}{a_i - a_j} \qquad (i,j = 1,2,\ldots,n;\ k,l = 1,2,\ldots,m;\ i \neq j).$$

Since K is assumed to be infinite, there is a $u \in K$ which is distinct from all these $(b_k - b_l)/(a_i - a_j)$. Hence

$$a_i u + b_l \neq a_j u + b_k \qquad \text{unless } i = j \text{ and } k = l. \qquad (*)$$

With this u, we put $c = au + b = a_1 u + b_1$. We claim $K(a,b) = K(c)$. Certainly $K(c) \subseteq K(a,b)$. In order to prove $K(a,b) \subseteq K(c)$, we must show $a,b \in K(c)$. Since $b = c - au$, the relation $a \in K(c)$ implies $b \in K(c)$. Hence we need only prove $a \in K(c)$. We do this by showing $x - a \in K(c)[x]$. We shall see that $x - a$ is a greatest common divisor of two polynomials in $K(c)[x]$.

Now $a = a_1$ is a root of $f(x)$ and of $g(c - ux)$. These are polynomials in $K(c)[x]$. Thus $x - a$ is a divisor of the greatest common divisor of $f(x)$ and $g(c - ux)$. On the other hand, any root a_i of $f(x)$ distinct from a_1 cannot be a root of $g(c - ux)$, because then $c - ua_i$ would be a root of $g(x)$, hence a_i would be equal to one of $b = b_1,b_2,\ldots,b_m$, contrary to $(*)$. Thus $a = a_1$ is the only common root of $f(x)$ and $g(c - ux)$. Thus $x - a$ is a greatest common divisor of the polynomials $f(x)$ and $g(c - ux)$ in $K(c)[x]$ and $x - a$ itself is in $K(c)[x]$. This gives $a \in K(c)$ and completes the proof. □

We can now prove that every finitely generated algebraic separable extension is a simple extension.

55.14 Theorem: *Let E/K be an algebraic separable extension of fields and assume $E = K(a_1,a_2, \ldots,a_m)$. Then there is an element c in E such that $E = K(c)$.*

Proof: We make induction on m. The claim is true when $m = 2$ by Theorem 55.13 (with $E = K(a,b)$). If the assertion is proved for $m - 1$, then $K(a_1,a_2, \ldots,a_{m-1}) = K(c_1)$ for some c_1 and therefore we have $K(a_1,a_2, \ldots,a_m) = K(a_1,a_2, \ldots,a_{m-1})(a_m) = K(c_1)(a_m) = K(c_1,a_m) = K(c)$ for some $c \in E$. □

We give a useful characterization of simple algebraic extensions. This yields an alternative proof of Theorem 55.14.

55.15 Theorem: *Let E/K be a finite-dimensional extension of fields. E is a simple extension of K if and only if there are only finitely many intermediate fields of E/K.*

Proof: Assume first that E is a simple extension of K, say $E = K(c)$. We want to show that there are finitely many intermediate fields. We will show that each intermediate field of E/K is uniquely determined by a divisor of the minimal polynomial of c over K.

Let $f(x) \in K[x]$ be the minimal polynomial of c over K. Let L be an intermediate field of E/K and let $g(x) \in L[x]$ be the minimal polynomial of c over L. The field L is generated over K by the coefficients of $g(x)$. To see this, let $g(x) = \sum_{i=0}^{m} a_i x^i$ (with $a_m = 1$) and $M = K(a_1,a_2, \ldots,a_m)$. Since $g(x)$ is in $L[x]$, we have $\{a_1,a_2, \ldots,a_m\} \subseteq L$ and $K(a_1,a_2, \ldots,a_m) \subseteq L$. Thus $M \subseteq L$ and $|E:M| \geqslant |E:L| = |K(c):L| = |L(c):L| = \deg g(x) = m$. On the other hand, c is a root of a polynomial $g(x)$ in $M[x]$ of degree m, so the degree of the minimal polynomial of c over M is at most m, so $|E:M| = |K(c):M| = |M(c):M| \leqslant m$ (Theorem 50.7). Therefore $|E:M| = m = |L(c):L| = |E:L|$ and consequently $|L:K| = |M:K|$. Together with $M \subseteq L$, this gives $M = L$ (Lemma 42:15(2)).

Therefore each intermediate field L of E/K is uniquely determined by the minimal polynomial $g(x)$ of the primitive element c over that intermediate field L. We know $g(x)$ divides $f(x)$ in $L[x]$ (Lemma 50.5). Let N

be a normal closure of E over K. Then N contains all roots of $f(x)$ and $f(x)$ splits in N. Of course $g(x)$ divides $f(x)$ in $N[x]$ and, since $N[x]$ is a unique factorization domain, $g(x)$ is a product of some of the linear factors of $f(x)$ in $N[x]$. Since, in $N[x]$, there are only finitely many monic divisors of $f(x)$, there is only a finite number of possibilities for $g(x)$ and there are only a finite number of intermediate fields L.

Assume conversely that there are only a finite number of intermediate fields of E/K. If K is finite, so is E and E is a simple extension of its prime subfield and of K (Theorem 52.19(1)). So we may suppose K is infinite. We choose an element c in E such that $|K(c):K|$ is as large as possible. In other words, $|K(c):K| \geqslant |K(b):K|$ for any $b \in E$. With this c, we claim $E = K(c)$. Otherwise, there is an $e \in E \setminus K(c)$. As k ranges through the infinite set K, we get finitely many intermediate fields $K(c + ek)$. Thus there are k and k' in K such that $k \neq k'$ and $K(c + ek) = K(c + ek')$. Then $c + ek$ and $c + ek'$ are in $K(c + ek)$, then their difference $e(k - k')$ is in $K(c + ek)$, then e is in $K(c + ek)$, hence ek is also in $K(c + ek)$ and finally $c = (c + ek) - ek$ is in $K(c + ek)$. Thus $e, c \in K(c + ek)$. So $K(c) \subseteq K(c + ek)$. Since $e \in K(c + ek)$ and $e \notin K(c)$, we get $K(c) \subset K(c + ek)$ and thus $|K(c):K| < |K(c + ek):K|$ (Lemma 42.15(2)), a contradiction. Hence $E = K(c)$. □

Theorem 55.14 follows very easily from Theorem 55.15. Suppose $E = K(a_1, a_2, \ldots, a_m)$ is an algebraic separable extension of K. We find a normal closure N of K over E. Then N is Galois over K and finite dimensional over K (Theorem 55.11). The Galois group of the extension N/K is thus finite and it has finitely many subgroups. By the fundamental theorem of Galois theory, there are finitely many intermediate fields of N/K and so finitely many intermediate fields of E/K. Theorem 55.15 states that E is a simple extension of K.

We proceed to prove the converse of Lemma 55.6. We need some prepatory lemmas, which are of intrinsic interest as well.

55.16 **Lemma:** *Let E/K be an extension of fields of characteristic $p \neq 0$ and let $a \in E$. Assume a is algebraic over K. Then a is separable over K if and only if $K(a) = K(a^p)$.*

Proof: Suppose first that a is separable over K. Then a is also separable over $K(a^p)$ by Lemma 55.5. Let $g(x) \in K(a^p)[x]$ be the minimal polynomial of a over $K(a^p)$. Thus all roots of $g(x)$ are simple. Since a is a root of the polynomial $x^p - a^p \in K(a^p)[x]$, we have $g(x)|x^p - a^p$ in $K(a^p)[x]$. Therefore $g(x)|x^p - a^p$ and $g(x)|(x - a)^p$ in $E[x]$. So $g(x) = (x - a)^m$ for some m such that $1 \leqslant m \leqslant p$. Since $g(x)$ has no multiple roots, we get $m = 1$. Then $g(x) = x - a \in K(a^p)[x]$ and consequently $a \in K(a^p)$. This gives $K(a) \subseteq K(a^p)$ and, since $K(a^p) \subseteq K(a)$ in any case, we obtain $K(a) = K(a^p)$.

Conversely, suppose that $K(a) = K(a^p)$. We want to show that a is separable over K. Let $f(x)$ be the minimal polynomial of a over K. If a is not separable over K, then $f(x)$ has the form $f(x) = g(x^p)$ for some $g(x) \in K[x]$. Here $g(x)$ is irreducible over K because $g(x)$ is not a unit in $K[x]$ (for $f(x)$, being irreducible over K, is not a unit in $K[x]$) and any factorization $g(x) = r(x)s(x)$ with $deg\ r(x) \neq 0 \neq deg\ s(x)$ would give a proper factorization $f(x) = r(x^p)s(x^p)$ with $deg\ r(x^p) \neq 0 \neq deg\ s(x^p)$, contrary to the irreducibility of $f(x)$ over K. Clearly $g(x)$ is a monic polynomial and, since $0 = f(a) = g(a^p)$, we see that a^p is a root of $g(x)$. Thus $g(x)$ is the minimal polynomial of a^p over K (Theorem 50.3). Of course $deg\ f(x) = p \cdot deg\ g(x)$ and

$$|K(a):K| = deg\ f(x) = p(deg\ g(x)) > deg\ g(x) = |K(a^p):K|.$$

Hence $K(a^p)$ is a *proper* subspace of the K-vector space $K(a)$ (Lemma 42.15(2)), contrary to the hypothesis $K(a) = K(a^p)$. Consequently, $K(a) = K(a^p)$ implies that a is separable over K. □

55.17 Lemma: *Let E/K be a finite dimensional extension of fields of characteristic $p \neq 0$, say $|E:K| = n$. Then the following are equivalent.*
(1) *There is a K-basis $\{u_1, u_2, \ldots, u_n\}$ of E such that $\{u_1^p, u_2^p, \ldots, u_n^p\}$ is also a K-basis of E.*
(2) *For all K-bases $\{t_1, t_2, \ldots, t_n\}$ of E, $\{t_1^p, t_2^p, \ldots, t_n^p\}$ is also a K-basis of E.*
(3) *E is a separable extension of K.*

Proof: (1) $\Longrightarrow$ (2) Let $\{u_1, u_2, \ldots, u_n\}$ be a such a K-basis of E that $\{u_1^p, u_2^p, \ldots, u_n^p\}$ is also a K-basis of E and let $\{t_1, t_2, \ldots, t_n\}$ be an arbitrary K-basis of E. In order to show that $\{t_1^p, t_2^p, \ldots, t_n^p\}$ is a K-basis of E, it suffices to prove that $\{t_1^p, t_2^p, \ldots, t_n^p\}$ spans E over K (Lemma 42.13(2); t_i^p are mutually distinct since $t_i^p - t_j^p = (t_i - t_j)^p \neq 0$ for $i \neq j$) and thus it

suffices to prove that $u_i^p \in s_K(t_1^p, t_2^p, \ldots, t_n^p)$ for all $i = 1,2,\ldots,n$. But this is obvious: we have

$$\begin{aligned} u_i &\in s_K(t_1, t_2, \ldots, t_n), \\ u_i &= k_1 t_1 + k_2 t_2 + \cdots + k_n t_n \qquad \text{for some } k_j \in K, \\ u_i^p &= k_1^p t_1^p + k_2^p t_2^p + \cdots + k_n^p t_n^p \qquad \text{for some } k_j^p \in K, \\ u_i^p &\in s_K(t_1^p, t_2^p, \ldots, t_n^p). \end{aligned}$$

(2) $\Longrightarrow$ (1) This is trivial.

(2) $\Longrightarrow$ (3) Suppose now that $\{t_1^p, t_2^p, \ldots, t_n^p\}$ is a K-basis of E whenever $\{t_1, t_2, \ldots, t_n\}$ is. Every element in E is algebraic over K because E/K is a finite dimensional extension (Theorem 50.10). Thus we are to show that every element b of E is separable over K. We do this by proving $K(b) = K(b^p)$ (Lemma 55.16).

Let $b \in E$. We put $r = |K(b):K|$. Then $r \leqslant n$ and $\{1,b,b^2, \ldots, b^{r-1}\}$ is a K-basis of $K(b)$. We extend the K-linearly independent subset $\{1,b,b^2, \ldots, b^{r-1}\}$ of E to a K-basis $\{1,b,b^2, \ldots, b^{r-1}, c_{r+1}, \ldots, c_n\}$ of E, as is possible by virtue of Theorem 42.14. Then $\{1,b^p,(b^p)^2, \ldots,(b^p)^{r-1}, c_{r+1}^p, \ldots, c_n^p\}$ is also a K-basis of E by hypothesis and so $\{1,b^p,(b^p)^2, \ldots,(b^p)^{r-1}\}$ is a K-linearly independent subset of $K(b)$. Lemma 42.13(1) states that $\{1,b^p,(b^p)^2, \ldots,(b^p)^{r-1}\}$ spans $K(b)$ over K. So $K(b) \subseteq s_K(1,b^p,(b^p)^2, \ldots,(b^p)^{r-1}) \subseteq K(b^p)$. This proves $K(b) = K(b^p)$. Hence b is separable over K.

(3) $\Longrightarrow$ (2) We assume E is separable over K and $\{t_1, t_2, \ldots, t_n\}$ is a K-basis of E. We want to show that $\{t_1^p, t_2^p, \ldots, t_n^p\}$ is a K-basis of E. Since $t_i^p \neq t_j^p$ for $i \neq j$, the set $\{t_1^p, t_2^p, \ldots, t_n^p\}$ has exactly $n = |E:K|$ elements and, in view of Lemma 42.13, it suffices to prove that $\{t_1^p, t_2^p, \ldots, t_n^p\}$ spans E over K. So we put $L = s_K(t_1^p, t_2^p, \ldots, t_n^p)$ and try to show $L = E$.

Our first step will be to establish that L is a subring of E. In order to prove this, we must only show that L is closed under multiplication. If $a = \sum_{i=1}^{n} a_i t_i^p$ and $b = \sum_{j=1}^{n} b_j t_j^p$ are elements of L ($a_i, b_j \in K$), then $ab = \sum_{i,j=1}^{n} a_i b_j t_i^p t_j^p$, and L will closed under multiplication provided $t_i^p t_j^p \in L$. As $\{t_1, t_2, \ldots, t_n\}$

is a K-basis of E, there are elements c_{ijk} in K with $t_i t_j = \sum_{i,j=1}^{n} c_{ijk} t_k$ and so $t_i^p t_j^p = \sum_{i,j=1}^{n} c_{ijk}^p t_k^p \in s_K(t_1^p, t_2^p, \ldots, t_n^p) = L$. Thus L is a subring of E.

Since L contains K and $\{t_1^p, t_2^p, \ldots, t_n^p\}$, and L is contained in the ring $K[t_1^p, t_2^p, \ldots, t_n^p]$, we get $L = K[t_1^p, t_2^p, \ldots, t_n^p]$. Now for each $i = 2, \ldots, n$, the element t_i^p is algebraic over K, so algebraic over $K(t_1^p, \ldots, t_{i-1}^p)$ and so $K(t_1^p, \ldots, t_{i-1}^p)[t_i] = K(t_1^p, \ldots, t_{i-1}^p)(t_i)$ (Theorem 50.6) and repeated application of Lemma 49.6(2), Lemma 49.6(3) gives $L = K[t_1^p, t_2^p, \ldots, t_n^p] = K(t_1^p, t_2^p, \ldots, t_n^p)$. Thus $L = K(t_1^p, t_2^p, \ldots, t_n^p)$ and L is in fact a field.

We now prove $E = K(t_1^p, t_2^p, \ldots, t_n^p)$. Let a be an arbitrary element of E. Then a is algebraic over K and over L (Lemma 50.5). Let $f(x) \in L[x]$ be the minimal polynomial of a over L. Since $a \in s_K(t_1, t_2, \ldots, t_n)$ and therefore $a^p \in s_K(t_1^p, t_2^p, \ldots, t_n^p) = L$, we see $x^p - a^p \in L[x]$ and a is a root of $x^p - a^p$. Thus $f(x)$ divides $x^p - a^p$ in $L[x]$. We put $x^p - a^p = f(x)^e g(x)$, where $e \geqslant 1$, $g(x) \in L[x] \setminus \{0\}$ and $(f(x), g(x)) \approx 1$ in $L[x]$. Taking derivatives, we obtain

$$0 = ef(x)^{e-1} f'(x) g(x) + f(x) g'(x),$$
$$g(x) \text{ divides } f(x) g'(x) \text{ in } L[x],$$
$$g(x) \text{ divides } g'(x) \text{ in } L[x],$$
$$g'(x) = 0,$$
$$0 = ef(x)^{e-1} f'(x) g(x),$$

and since E is separable over K, here $f'(x) \neq 0$, so $f(x)^{e-1} f'(x) g(x) \neq 0$ and

$$e = 0 \text{ in } L,$$
$$p \mid e \quad \text{in } \mathbb{Z},$$
$$e = pm \text{ for some } m \in \mathbb{N},$$
$$p = deg(x^p - a^p) = pm(deg\ f(x)) + deg\ g(x),$$
$$m = 1,\ deg\ f(x) = 1 \text{ and } g(x) = 0,$$
$$e = p \text{ and } g(x) = 1 \text{ (comparing leading coefficients)},$$
$$(x - a)^p = x^p - a^p = f(x)^p,$$
$$x - a = f(x) \in L[x],$$

and $a \in L$. This proves $E \subseteq L$. Hence $E = L = s_K(t_1^p, t_2^p, \ldots, t_n^p)$ and thus $\{t_1^p, t_2^p, \ldots, t_n^p\}$ is a K-basis of E, as was to be proved. □

55.18 Lemma: *Let E/K be a field extension and $a \in E$. Then $K(a)$ is a separable extension of K if and only if a is separable over K.*

Proof: If $K(a)$ is separable over K, then every element of $K(a)$ is separable over K, in particular a is separable over K. Suppose now a is separable (thus algebraic) over K. We wish to prove that $K(a)$ is separable over K. The case *char* $K = 0$ being trivial, we may assume *char* $K = p \neq 0$. Let $n = |K(a):K|$. Then $\{1,a,a^2, \ldots, a^{n-1}\}$ is a K-basis of $K(a)$ (Theorem 50.7). Likewise $\{1,a^p,(a^p)^2, \ldots,(a^p)^{m-1}\}$ is a K-basis of $K(a^p)$, where $m = |K(a^p):K|$. Since a is separable over K, we have $K(a^p) = K(a)$ (Lemma 55.16) and $m = |K(a^p):K| = |K(a):K| = n$. Thus $\{1,a^p,(a^p)^2, \ldots,(a^p)^{n-1}\} = \{1^p,(a)^p,(a^2)^p, \ldots,(a^{n-1})^p\}$ is also a K-basis of $K(a)$. Thus $K(a)$ is separable over K by Lemma 55.17. □

55.19 Theorem: *Let E/K be a finite dimensional field extension and let L be an intermediate field of E/K. Then E is separable over K if and only if E is separable over L and L is separable over K.*

Proof: If E is separable over K, then E is separable over L and L is separable over K (Lemma 55.6). Conversely, suppose that E is separable over L and L is separable over K. We are to show that (1) E is algebraic over K and that (2) any element in E is separable over K. Since E/L and L/K are separable extensions, they are algebraic extensions and E/K is also algebraic by Theorem 50.16. Now the separability of E over K. The case *char* $K = 0$ being trivial, we assume *char* $K = p \neq 0$. As E/K is a finite dimensional extension by hypothesis, $|E:L|$ and $|L:K|$ are finite (Lemma 48.14). Let $|E:L| = n$ and $|L:K| = m$.

Since E is separable over L, there is an L-basis $\{a_1,a_2, \ldots,a_n\}$ of E such that $\{a_1^p,a_2^p, \ldots,a_n^p\}$ is also an L-basis of E and, since L is separable over K, there is a K-basis $\{b_1,b_2, \ldots,b_m\}$ of L such that $\{b_1^p,b_2^p, \ldots,b_m^p\}$ is also a K-basis of L (Lemma 55.17). Then $\{a_ib_j\}$ is a K-basis of E by the proof of Theorem 48.13, and likewise $\{a_i^p b_j^p\}$ is a K-basis of E. Hence $\{a_ib_j\}$ is a K-basis of E such that $\{(a_ib_j)^p\}$ is also a K-basis of E. From Lemma 55.17, it follows that E is separable over K. □

We close this paragraph with a brief discussion of perfect fields.

55.20 Definition: Let K be a field. If *char* $K = 0$ or if *char* $K = p \neq 0$ and for each $a \in K$, there is a $b \in K$ such that $a = b^p$, then K is said to be *perfect*.

Thus in case *char* $K = p \neq 0$, K is a perfect field if and only if the field homomorphism $\varphi: K \to K$ $(u \to u^p)$ is *onto* K. Then for each $a \in K$, there is a *unique* $b \in K$ such that $a = b^p$, for φ is one-to-one. This unique b will be denoted by $\sqrt[p]{a}$.

For example, every finite field is perfect, for if $\mathbb{F}_q$ is a finite field and *char* $\mathbb{F}_q = p \neq 0$, then the one-to-one homomorphism $\varphi: \mathbb{F}_q \to \mathbb{F}_q$ $(u \to u^p)$ is $\mathbb{F}_p$-linear and thus onto $\mathbb{F}_q$ by Theorem 42.22 (or, more simply, because the one-to-one mapping from the finite set $\mathbb{F}_q$ into $\mathbb{F}_q$ must be *onto* $\mathbb{F}_q$).

55.21 Theorem: *Let K be a field. K is perfect if and only if every irreducible polynomial in $K[x]$ is separable over K.*

Proof: The assertion is trivial in case *char* $K = 0$, so assume that *char* $K = p \neq 0$.

Suppose first that K is perfect. Now, if $f(x) \in K[x]$ is not separable over K, then $f(x) = g(x^p)$ for some $g(x) \in K[x]$, say $g(x) = \sum_{i=0}^{m} a_i x^i$ and

$$f(x) = g(x^p) = \sum_{i=0}^{m} a_i x^{ip} = \sum_{i=0}^{m} (\sqrt[p]{a_i})^p x^{ip} = (\sum_{i=0}^{m} \sqrt[p]{a_i}\, x^i)^p,$$

$f(x)$ cannot be irreducible over K. Thus, if K is a perfect field, then every irreducible polynomial in $K[x]$ is separable over K.

Conversely, suppose that every irreducible polynomial in $K[x]$ is separable over K. We want to show that K is perfect. Let $a \in K$. We must find a b in K with $b^p = a$. So we consider the polynomial $x^p - a \in K[x]$. We adjoin a root $\sqrt[p]{a}$ of $x^p - a$ to K and obtain the field $K(\sqrt[p]{a})$ (possibly $K \subset K(\sqrt[p]{a})$). Then, in $K(\sqrt[p]{a})[x]$, we have the factorization $x^p - a = (x - \sqrt[p]{a})^p$. The mini-

mal polynomial of $\sqrt[p]{a}$ over K is thus $(x - \sqrt[p]{a})^k$ for some $k \in \{1,2,\ldots,p\}$. So $(x - \sqrt[p]{a})^k$ is necessarily irreducible and, by hypothesis, separable over K and has therefore no multiple roots. This forces $k = 1$. So the minimal polynomial of $\sqrt[p]{a}$ is $x - \sqrt[p]{a} \in K[x]$, which gives $\sqrt[p]{a} \in K$, as was to be proved. □

Consequently every algebraic extension of a perfect field K is separable over K. Theorem 55.21 yields the corollary that every algebraically closed field is perfect, since any irreducible polynomial in an algebraically closed field is of first degree and has therefore no multiple roots (is separable over that field).

Exercises

1. Find a normal closure of $\mathbb{Q}(\sqrt{3},\sqrt[5]{7})$ over $\mathbb{Q}$.

2. If E/K is a field extension and $|E:K| = 2$, show that E is normal over K.

3. Let E/K be a field extension with $|E:K| = 3$ and assume that E is not normal over K. Let N be a normal closure of E over K. Show that $|N:K| = 6$ and that there is a unique intermediate field L of N/K satisfying $|L:K| = 2$.

4. Let N/K be a field extension and assume that N is normal over K. Let L be an intermediate field of N/K. Prove that L is normal over K if and only if E is (K,N)-stable.

5. Find fields $K \subseteq L \subseteq N$ such that N is normal over L, L is normal over K but N is not normal over K.

6. Find fields $K \subseteq L \subseteq N$ such that $|N:K| = 6$, N is Galois over K but L is not Galois over K.

7. Find fields $K \subseteq L \subseteq N$ such that $|N:K|$ is finite, N is normal over K but L is not normal over K.

8. Find primitive elements for the extensions $\mathbb{Q}(\sqrt{2},\sqrt{3})$, $\mathbb{Q}(\sqrt{2},\sqrt{3},\sqrt{5})$, $\mathbb{Q}(\sqrt{2},i)$, $\mathbb{Q}(\sqrt{2},\sqrt[3]{5})$ of $\mathbb{Q}$.

9. Find a splitting field K over $\mathbb{F}_3$ of $(x^2 + 1)(x^2 + x + 2) \in \mathbb{F}_3[x]$ and a primitive element of K.

10. Let p be a prime number and x,y two distinct indeterminates over $\mathbb{F}_p$. Let $E = \mathbb{F}_p(x,y)$ and $K = \mathbb{F}_p(x^p,y^p)$. Show that E is not a simple extension of K and find infinitely many intermediate fields of E/K.

11. Prove the following generalization of Theorem 55.14. If K is a field, $K(a_1,a_2,\ldots,a_m)$ is an algebraic extension of K and $a_2,\ldots,a_m$ are separable over K, then $K(a_1,a_2,\ldots,a_m)$ is a simple extension of K.

12. Let K be a field and $K(a_1,a_2,\ldots,a_m)$ a finitely generated extension of K. Show that $K(a_1,a_2,\ldots,a_m)$ is separable over K if and only if all $a_1,a_2,\ldots,a_m$ are separable over K.

13. Prove that Theorem 55.19 is valid without the hypothesis that E be finite dimensional over K. (Hint: Reduce the general case to the finite dimensional case.)

14. Let L and M be intermediate fields of a field extension E/K. Prove that, if L is separable over K, then LM is separable over M.

15. Prove that every finite dimensional extension of a perfect field is perfect.

16. Let E/K be a finite dimensional field extension. If E is perfect, show that K is also perfect.

§56
Galois Group of a Polynomial

In this paragraph, we give some applications of Galois theory to the theory of equations. We shall introduce resultants and discriminants, and then discuss polynomial equations $f(x) = 0$, where $f(x)$ is of degree 2,3,4.

56.1 Lemma: *Let K be a field and $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, $g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$ be nonzero polynomials in $K[x]\backslash K$. Assume that at least one of a_n, b_m is distinct from 0. Then $f(x), g(x)$ have a nonunit greatest common divisor in $K[x]$ if and only if there are nonzero polynomials $g_1(x), f_1(x) \in K[x]$ such that*

$$f(x)g_1(x) = g(x)f_1(x) \quad \text{and} \quad \deg f_1(x) < n, \deg g_1(x) < m.$$

Proof: One direction is clear. If $f(x)$ and $g(x)$ have a nonunit greatest common divisor $h(x)$ in $K[x]$, then $f(x) = h(x)f_1(x)$, $g(x) = h(x)g_1(x)$ with some suitable $f_1(x), g_1(x)$ in $K[x]$ and

$$\deg f_1(x) = \deg f(x) - \deg h(x) \leqslant n - \deg h(x) < n$$

since $\deg h(x)$ is greater than zero. Likewise $\deg g_1(x) < m$. We have of course $f(x)g_1(x) = f_1(x)h(x)g_1(x) = f_1(x)g(x)$.

Conversely, assume $f(x)g_1(x) = g(x)f_1(x)$ for some nonzero polynomials $f_1(x)$, $g_1(x)$ in $K[x]$ satisfying $\deg f_1(x) < n$ and $\deg g_1(x) < m$. We put $h(x) \approx (f(x), g(x))$. We want to prove $\deg h(x) > 0$. Write $f(x) = h(x)F(x)$, $g(x) = h(x)G(x)$. Then $(F(x), G(x)) \approx 1$ and $f(x)g_1(x) = g(x)f_1(x)$ gives $F(x)g_1(x) = G(x)f_1(x)$. Suppose, without loss of generality, $a_n \neq 0$, so that $\deg f(x) = n$. Now $F(x)$ divides $G(x)f_1(x)$ and, as $(F(x), G(x)) \approx 1$, $F(x)$ divides $f_1(x)$; thus $\deg F(x) \leqslant \deg f_1(x) < n = \deg f(x) = \deg F(x) + \deg h(x)$ and we get $\deg h(x) > 0$. This completes the proof. □

Let K be a field and

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$
$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$$

two polynomials in $K[x]$, where $a_n \neq 0$ or $b_m \neq 0$, so that $deg\ f(x) = n$ or $deg\ g(x) = m$. From Lemma 56.1, we know that $f(x)$ and $g(x)$ have a nonunit greatest common divisor in $K[x]$ if and only if there are elements $c_{m-1}, c_{m-2}, \ldots, c_1, c_0, d_{n-1}, d_{n-2}, \ldots, d_1, d_0$, where at least one $c_i \neq 0$ and at least one $d_j \neq 0$, such that

$$(a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0)(c_{m-1}x^{m-1} + c_{m-2}x^{m-2} + \cdots + c_1 x + c_0)$$
$$= (b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0)(d_{n-1}x^{n-1} + d_{n-2}x^{n-2} + \cdots + d_1 x + d_0). \quad (*)$$

This polynomial equation is equivalent to the system of equations:

$$\begin{aligned}
a_n c_{m-1} &= b_m d_{n-1} \\
a_n c_{m-2} + a_{n-1}c_{m-1} &= b_m d_{n-2} + b_{m-1}d_{n-1} \\
a_n c_{m-3} + a_{n-1}c_{m-2} + a_{n-2}c_{m-1} &= b_m d_{n-3} + b_{m-1}d_{n-2} + b_{m-2}d_{n-1} \\
&\cdots\cdots\cdots\cdots \\
a_1 c_0 + a_0 c_1 &= b_1 d_0 + b_0 d_1 \\
a_0 c_0 &= b_0 d_0.
\end{aligned}$$

This system can be written as

$$\begin{aligned}
a_n c_{m-1} \qquad\qquad - b_m d_{n-1} \qquad\qquad &= 0 \\
a_n c_{m-2} + a_{n-1}c_{m-1} \qquad - b_m d_{n-2} - b_{m-1}d_{n-1} \qquad &= 0 \\
a_n c_{m-3} + a_{n-1}c_{m-2} + a_{n-2}c_{m-1} \quad - b_m d_{n-3} - b_{m-1}d_{n-2} - b_{m-2}d_{n-1} \quad &= 0 \\
&\cdots\cdots\cdots\cdots \\
a_1 c_0 + a_0 c_1 \qquad\qquad - b_1 d_0 - b_0 d_1 \quad &= 0 \\
a_0 c_0 \qquad\qquad - b_0 d_0 \quad &= 0
\end{aligned}$$

or as

$$\begin{aligned}
a_n c_{m-1} \qquad\qquad -b_m d_{n-1} \qquad\qquad &= 0 \\
a_{n-1}c_{m-1} + a_n c_{m-2} \qquad -b_{m-1}d_{n-1} - b_m d_{n-2} \qquad &= 0 \\
a_{n-2}c_{m-1} + a_{n-1}c_{m-2} + a_n c_{m-3} \quad -b_{m-2}d_{n-1} - b_{m-1}d_{n-2} - b_m d_{n-3} \quad &= 0 \\
&\cdots\cdots\cdots\cdots \\
a_1 c_{m-1} + a_2 c_{m-2} + a_3 c_{m-3} \qquad \cdots\cdots\cdots \qquad &= 0 \\
a_0 c_{m-1} + a_1 c_{m-2} + a_2 c_{m-3} \qquad \cdots\cdots\cdots \qquad &= 0 \\
a_0 c_{m-2} + a_1 c_{m-3} \qquad \cdots\cdots\cdots \qquad &= 0 \\
&\cdots\cdots\cdots\cdots \\
a_0 c_1 + a_1 c_0 \qquad\qquad - b_0 d_1 - b_1 d_0 \quad &= 0 \\
a_0 c_0 \qquad\qquad - b_0 d_0 \quad &= 0
\end{aligned}$$

We write this system in matrix form:

$$\begin{array}{c} \overbrace{}^{m \text{ columns}} \quad \overbrace{}^{n \text{ columns}} \\ \begin{pmatrix} a_n & 0 & 0 & \dots & 0 & b_m & 0 & 0 & \dots & 0 \\ a_{n-1} & a_n & 0 & \dots & 0 & b_{m-1} & b_m & 0 & \dots & 0 \\ a_{n-2} & a_{n-1} & a_n & \dots & 0 & b_{m-2} & b_{m-1} & b_m & \dots & 0 \\ \multicolumn{10}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots} \\ 0 & 0 & 0 & \dots & a_0 & 0 & 0 & 0 & \dots & b_0 \end{pmatrix} \end{array} \begin{pmatrix} c_m \\ c_{m-2} \\ c_{m-3} \\ \vdots \\ -d_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Let A denote the matrix of this system. Then the polynomials $f(x),g(x)$ have a nonunit greatest common divisor if and only if the matrix equation $AX = 0$ has a solution

$$X = (c_{m-1}, c_{m-2}, c_{m-3}, \dots, c_1, c_0, -d_{n-1}, -d_{n-2}, -d_{n-3}, \dots, -d_1, -d_0)^t$$

in which at least one $c_i \neq 0$ and at least one $d_j \neq 0$. From the equation (*) and the fact that $K[x]$ has no zero divisors, we deduce that, in a solution $X = (c_{m-1}, \dots, -d_0)^t$ of $AX = 0$, there is at least one $c_i \neq 0$ if and only if there is at least one $d_j \neq 0$. Thus the polynomials $f(x),g(x)$ have a nonunit greatest common divisor if and only if the matrix equation $AX = 0$ has a nontrivial solution. This is the case if and only if $det\ A = 0$ (Theorem 45.3). Since $det\ A = det\ A^t$, we get that $f(x),g(x)$ have a nonunit greatest common divisor if and only if $det\ A^t = 0$. We proved the

56.2 Theorem: *Let K be a field and $f(x) = a_n x^n + a_{n-1}x^{n-1} + \dots + a_1 x + a_0$, $g(x) = b_m x^m + b_{m-1}x^{m-1} + \dots + b_1 x + b_0$ be polynomials in $K[x]\backslash K$, where at least one of a_n, b_m is distinct from 0. Then $f(x)$ and $g(x)$ have a nonunit greatest common divisor in $K[x]$ if and only if the determinant*

$$\begin{vmatrix}
a_n & a_{n-1} & \cdots & a_1 & a_0 & 0 & 0 & 0 & 0 & 0 \\
0 & a_n & a_{n-1} & \cdots & a_1 & a_0 & 0 & 0 & 0 & 0 \\
0 & 0 & a_n & a_{n-1} & \cdots & a_1 & a_0 & 0 & 0 & 0 \\
 & & & \cdots\cdots\cdots & & & & & & \\
0 & 0 & 0 & \cdots & & a_n & a_{n-1} & \cdots & a_1 & a_0 \\
b_m & b_{m-1} & \cdots & b_1 & b_0 & 0 & 0 & 0 & 0 & 0 \\
0 & b_m & b_{m-1} & \cdots & b_1 & b_0 & 0 & 0 & 0 & 0 \\
0 & 0 & b_m & b_{m-1} & \cdots & b_1 & b_0 & 0 & 0 & 0 \\
 & & & \cdots\cdots\cdots & & & & & & \\
0 & 0 & 0 & \cdots & & b_m & b_{m-1} & \cdots & b_1 & b_0
\end{vmatrix}$$

is equal to zero. □

56.3 Definition: Let K be a field and $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, $g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$ polynomials in $K[x]$. The determinant

$$\begin{vmatrix}
a_n & a_{n-1} & \cdots & a_1 & a_0 & & & \\
 & a_n & a_{n-1} & \cdots & a_1 & a_0 & & \\
 & & a_n & a_{n-1} & \cdots & a_1 & a_0 & \\
 & & & \cdots\cdots\cdots & & & & \\
 & & & & a_n & a_{n-1} & \cdots & a_1 & a_0 \\
b_m & b_{m-1} & \cdots & b_1 & b_0 & & & & \\
 & b_m & b_{m-1} & \cdots & b_1 & b_0 & & & \\
 & & b_m & b_{m-1} & \cdots & b_1 & b_0 & & \\
 & & & \cdots\cdots\cdots & & & & & \\
 & & & & b_m & b_{m-1} & \cdots & b_1 & b_0
\end{vmatrix}$$

(empty places are to be filled with zeroes) is called the *resultant of* $f(x)$ *and* $g(x)$, and is denoted by $R(f,g)$ or by $R(f(x),g(x))$.

56.4 Remark: Notice that a_n and b_m can be zero in Definition 56.3. There is ambiguity in this definition and notation: the resultant depends not only on $f(x)$ and $g(x)$, but also on the number of apparent coefficients, a point neglected in almost every book. For example, let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ and $g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$

again, and let $b_{m+1} = 0$, $h(x) = b_{m+1}x^{m+1} + b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$. Then of course $g(x) = h(x)$ but $R(f,h)$ has one more column than $R(f,g)$ and the expansion of $R(f,h)$ along the first column gives $R(f,h) = a_n R(f,g)$, so $R(f,h) \neq R(f,g)$ (unless $a_n = 1$ or $R(f,g) = 0$). Thus adding an initial term to $g(x)$ with coefficient 0 changes $R(f,g)$ to $a_n R(f,g)$. Consequently, if

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

$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$ and $b_m = b_{m-1} = \cdots = b_{k+1} = 0$, $b_k \neq 0$,

$G(x) = b_k x^k + b_{k-1}x^{k-1} + \cdots + b_1 x + b_0$,

then $g(x)$ is obtained from $G(x)$ by adding $m - k$ initial terms $b_m x^m$, $b_{m-1}x^{m-1}, \ldots, b_{k+1}x^k$ with coefficient 0 and so $R(f,g) = a_n^{m-k} R(f,G)$.

Definition 56.3 gives a new formulation of Theorem 56.2

56.2 Theorem: *Let K be a field and $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, $g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$ be polynomials in $K[x]\backslash K$, where at least one of a_n, b_m is distinct from 0. Then $f(x)$ and $g(x)$ have a nonunit greatest common divisor in $K[x]$ if and only if $R(f,g) = 0$.*

We give some product formulas for the resultant of two polynomials. These formulas make it evident that the resultant is 0 if and only if the polynomials have a nontrivial common factor.

56.5 Theorem: *Let K be a field and $u_1, u_2, \ldots, u_n, y_1, y_2, \ldots, y_m$, indeterminates over K. Let a_n, b_m be nonzero elements of K and let x be an indeterminate over K distinct from all of $u_1, u_2, \ldots, u_n, y_1, y_2, \ldots, y_m$. Let $f(x)$ and $g(x)$ be polynomials in $K(u_1, u_2, \ldots, u_n, y_1, y_2, \ldots, y_m)[x]$ defined by*

$$f(x) = a_n(x - u_1)(x - u_2)\ldots(x - u_n)$$
$$g(x) = b_m(x - y_1)(x - y_2)\ldots(x - y_m).$$

Then the following hold.

(1) *$R(f,g)$ is in $P[a_n, u_1, u_2, \ldots, u_n, b_m, y_1, y_2, \ldots, y_m]$, where P is the prime subfield of K.*

(2) $R(f,g) = a_n^m b_m^n \prod_{i=1}^{n} \prod_{j=1}^{m} (u_i - y_j)$.

(3) $R(f,g) = a_n^m \prod_{i=1}^{n} g(u_i)$.

(4) $R(f,g) = (-1)^{mn} b_m^n \prod_{j=1}^{m} f(y_j)$.

Proof: We put

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$
$$g(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0,$$

where $a_i, b_j \in K(u_1, u_2, \ldots, u_n, y_1, y_2, \ldots, y_m)$. Thus $R(f,g)$ is a determinant of a matrix whose entries are $a_0, a_1, a_2, \ldots, a_n, b_0, b_1, b_2, \ldots, b_m$ and 0. Hence the entries of the matrix are in $P[a_0, a_1, a_2, \ldots, a_n, b_0, b_1, b_2, \ldots, b_m]$ and the determinant $R(f,g)$ itself is also in $P[a_0, a_1, a_2, \ldots, a_n, b_0, b_1, b_2, \ldots, b_m]$ (Remark 44.2(2)). Since each a_i/a_n, aside from a sign, is an elementary symmetric polynomial in $u_1, u_2, \ldots, u_n$, and since the coefficients of elementary symmetric polynomials are in the prime subfield P, we get

$$a_i/a_n \in P[u_1, u_2, \ldots, u_n] \text{ for all } i = 1,2, \ldots, n.$$

So each a_i is in $P[a_n, u_1, u_2, \ldots, u_n] \subseteq P[a_n, u_1, u_2, \ldots, u_n, b_m, y_1, y_2, \ldots, y_m]$. Likewise each b_j is in $P[a_n, u_1, u_2, \ldots, u_n, b_m, y_1, y_2, \ldots, y_m]$. Consequently

$$R(f,g) \in P[a_0, a_1, a_2, \ldots, a_n, b_0, b_1, b_2, \ldots, b_m] \subseteq P[a_n, u_1, u_2, \ldots, u_n, b_m, y_1, y_2, \ldots, y_m].$$

This proves (1). Now let $L = P[a_n, u_1, u_2, \ldots, u_n, b_m, y_1, y_2, \ldots, y_m]$. We put

$$S = a_n^m b_m^n \prod_{i=1}^{n} \prod_{j=1}^{m} (u_i - y_j) \in L.$$

We have

$$g(x) = b_m \prod_{j=1}^{m} (x - y_j),$$

$$g(u_i) = b_m \prod_{j=1}^{m} (u_i - y_j),$$

$$\prod_{i=1}^{n} g(u_i) = b_m^n \prod_{i=1}^{n} \prod_{j=1}^{m} (u_i - y_j).$$

and thus

$$S = a_n^m \prod_{i=1}^{n} g(u_i). \qquad \text{(i)}$$

In like manner, from $f(x) = a_n \prod_{i=1}^{n} (x - u_i) = (-1)^n a_n \prod_{i=1}^{n} (u_i - x)$, we get

$$f(y_j) = (-1)^n a_n \prod_{i=1}^{n} (u_i - y_j),$$

$$\prod_{j=1}^{m} f(y_j) = \prod_{j=1}^{m} \left((-1)^n a_n \prod_{i=1}^{n} (u_i - y_j)\right),$$

$$\prod_{j=1}^{m} f(y_j) = (-1)^{nm} a_n^m \prod_{j=1}^{m} \prod_{i=1}^{n} (u_i - y_j),$$

$$S = (-1)^{nm} b_m^n \prod_{j=1}^{m} f(y_j). \qquad \text{(ii)}$$

Now let $f_0(x)$ be the polynomial obtained by substituting y_j for u_i in $f(x)$. Thus

$$f_0(x) = a_n(x - u_1)\ldots(x - u_{i-1})(x - y_j)(x - u_{i+1})(x - u_n)$$
$$\in P(a_n, u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n, b_m, y_1, y_2, \ldots, y_m)[x].$$

Then the polynomials $f_0(x)$ and $g(x)$ in

$$P(a_n, u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n, b_m, y_1, y_2, \ldots, y_m)[x]$$

have a common factor $x - y_j$ and therefore $R(f_0, g) = 0$.

Thus $R(f,g) \in L$, regarded as a polynomial in

$$P[a_n, u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n, b_m, y_1, y_2, \ldots, y_m][u_i]$$

has the value $R(f_0, g) = 0$ when y_j is substituted for u_i. So $R(f,g)$ has the root y_j. So $u_i - y_j$ divides $R(f,g)$ in

$$P[a_n, u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n, b_m, y_1, y_2, \ldots, y_m][u_i] = L.$$

This is true for all $i = 1,2, \ldots, n$ and for all $j = 1,2, \ldots, m$. Since any $u_i - y_j$ is irreducible in L, and $u_i - y_j$ is distinct from $u_{i'} - y_{j'}$ whenever $(i,j) \neq (i',j')$, the polynomials $u_i - y_j$ are pairwise relatively prime. Thus $R(f,g)$ is divisible, in L, by their product

$$\prod_{i=1}^{n} \prod_{j=1}^{m} (u_i - y_j).$$

It follows that $R(f,g)$ is divisible by

$$S = a_n^m b_m^n \prod_{i=1}^{n} \prod_{j=1}^{m} (u_i - y_j)$$

in $M[u_1,u_2, \ldots,u_n,y_1,y_2, \ldots,y_m]$, where we put $M = P(a_n,b_m)$.

Let us write $H = R(f,g)/S$. Basically, we will argue that $R(f,g)$ and S are both homogeneous (§35, Ex. 4) of the same degree and conclude that H is a constant. Comparison of a monomial appearing in these polynomials will yield that this constant must be equal to 1, whence $R(f,g) = S$. The details are rather tedious.

From (i), we see that $S/a_n^m = \prod_{i=1}^{n} g(u_i)$

$$\begin{aligned} &= (b_m u_1^m + \cdots)(b_m u_2^m + \cdots)\ldots(b_m u_n^m + \cdots) \\ &= b_m^n u_1^m u_2^m \ldots u_n^m + \cdots \\ &\in P[b_m,y_1,y_2, \ldots,y_m][u_1,u_2, \ldots,u_n] \end{aligned}$$

is a symmetric polynomial in $u_1,u_2, \ldots,u_n$ over $P[b_m,y_1,y_2, \ldots,y_m]$ and hence there is a unique polynomial h_1 in n indeterminates over the integral domain $P[b_m,y_1,y_2, \ldots,y_m]$ such that

$$S/a_n^m = h_1(-a_{n-1}/a_n, a_{n-2}/a_n, \ldots, \mp a_1/a_n, \pm a_0/a_n).$$

Let us recall that h_1 is obtained from S/a_n^m by subtracting symmetric polynomials of the form

$$y\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{n-1}^{k_{n-1}-k_n}\sigma_n^{k_n}, \qquad y \in P[b_m,y_1,y_2, \ldots,y_m]$$

where $y u_1^{k_1}u_2^{k_2}\ldots u_{n-1}^{k_{n-1}}u_n^{k_n}$ are certain monomials appearing in S/a_n^m. We have $m \geqslant k_1$ by Lemma 38.8(2) since the leading monomial of S/a_n^m is $b_m^n u_1^m u_2^m \ldots u_n^m$. A symmetric polynomial of the form above gives rise to a term

$$y(-a_{n-1}/a_n)^{k_1-k_2}(a_{n-2}/a_n)^{k_2-k_3}\ldots(\mp a_1/a_n)^{k_{n-1}-k_n}(\pm a_0/a_n)^{k_n},$$

which is $(1/a_n)^{k_1}$ times a polynomial in $P[b_m,y_1,y_2, \ldots,y_m][a_0,a_1, \ldots,a_{n-1}]$. As $m \geqslant k_1$ for each of the terms in h_1, we see $a_n^m h_1$ is a polynomial in $P[b_m,y_1,y_2, \ldots,y_m][a_0,a_1, \ldots,a_{n-1},a_n]$. Thus

$$S = (a_n^m)(S/a_n^m) = a_n^m h_1(-a_{n-1}/a_n, a_{n-2}/a_n, \ldots, \mp a_1/a_n, \pm a_0/a_n),$$

$$S \in P[b_m,y_1,y_2,\ldots,y_m][a_0,a_1,\ldots,a_{n-1},a_n] \qquad \text{(iii)}$$

and $S = h(a_0,a_1,\ldots,a_{n-1},a_n)$, where h is a polynomial in $n+1$ indeterminates over $P[b_m,y_1,y_2,\ldots,y_m]$ (Lemma 49.5(1)).

Also
$$\begin{aligned} R(f,g) &\in P[b_0,b_1,b_2,\ldots,b_m][a_0,a_1,a_2,\ldots,a_n] \\ &\subseteq P[b_m,y_1,y_2,\ldots,y_m][a_0,a_1,\ldots,a_{n-1},a_n] \end{aligned}$$

and, together with (iii), we obtain

$$H = R(f,g)/S \in P[b_m,y_1,y_2,\ldots,y_m](a_0,a_1,\ldots,a_{n-1},a_n).$$

Thus $H \in M[y_1,y_2,\ldots,y_m][u_1,u_2,\ldots,u_n]$ is symmetric in $u_1,u_2,\ldots,u_n$ and therefore

$$H = k(-a_{n-1}/a_n, a_{n-2}/a_n, \ldots, \mp a_1/a_n, \pm a_0/a_n)$$

for some polynomial k in n indeterminates over $M[b_m,y_1,y_2,\ldots,y_m]$, which gives $H \in M[b_m,y_1,y_2,\ldots,y_m][a_0,a_1,\ldots,a_{n-1},a_n]$ (Lemma 49.5(1)).

Now $H = R(f,g)/S = R(f,g)/h(a_0,a_1,\ldots,a_{n-1},a_n)$. Note that multiplying the coefficients $a_n,a_{n-1},\ldots,a_1,a_0$ of $f(x)$ by an indeterminate t does not change the roots $u_1,u_2,\ldots,u_n$ of $f(x)$, but, in view of (i), changes S to t^mS, so that

$$h(ta_n,ta_{n-1},\ldots,ta_1,ta_0) = t^m h(a_n,a_{n-1},\ldots,a_1,a_0).$$

Likewise multiplying the coefficients $a_n,a_{n-1},\ldots,a_1,a_0$ of $f(x)$ by an indeterminate t changes $R(f,g)$ to $t^mR(f,g)$, as the determinant $R(f,g)$ has m rows consisting of zeroes and the coefficients of f. Thus H does not change when the coefficients of f are multiplied by t. But any monomial

$$y{a_0}^{k_0}{a_1}^{k_1}\ldots{a_n}^{k_n} \qquad (y \in M[b_m,y_1,y_2,\ldots,y_m])$$

changes then to $y(ta_0)^{k_0}(ta_1)^{k_1}\ldots(ta_n)^{k_n} = t^{k_0+k_1+\cdots+k_n}y{a_0}^{k_0}{a_1}^{k_1}\ldots{a_n}^{k_n}$. Thus the exponent system of any monomial $y{a_0}^{k_0}{a_1}^{k_1}\ldots{a_n}^{k_n}$ appearing in H is such that $k_0 + k_1 + \cdots + k_n = 0$. This means $k_0 = k_1 = \cdots = k_n = 0$ for all monomials $y{a_0}^{k_0}{a_1}^{k_1}\ldots{a_n}^{k_n}$ appearing in H and H is a "constant", i.e., H is in $M[b_m,y_1,y_2,\ldots,y_m]$.

Repeating the same argument with S/b_m^n in place of S/a_n^m, we get that H is in $M[a_n,u_1,u_2,\ldots,u_n]$. So $H \in M = P(a_n,b_m) \subseteq K$.

Thus $R(f,g) = HS$ for some $H \in K$. The constant term in $S = a_n^m \prod_{i=1}^{n} g(u_i)$ is equal to $a_n^m b_0^n$. So $R(f,g)$ must have a term $H a_n^m b_0^n$. Now $R(f,g)$ has the term $a_n^m b_0^n$, the product of the entries in the principal diagonal. Hence $H = 1$ and $R(f,g) = S$. This proves (2). From (i) and (ii), we get the equations in (3) and (4). □

56.6 Lemma: *Let K be a field and $f(x), g(x)$ polynomials of positive degree in $K[x]$, say deg $f(x) = n$ and deg $g(x) = m$. Let a_n be the leading coefficient of $f(x)$ and b_m the leading coefficient of $g(x)$. Let $r_1, r_2, \ldots, r_n$ be roots of $f(x)$ and $s_1, s_2, \ldots, s_m$ roots of $g(x)$ in a splitting field of $f(x)g(x)$ over K. Then*

$$R(f,g) = a_n^m b_m^n \prod_{i=1}^{n} \prod_{j=1}^{m} (r_i - s_j) = a_n^m \prod_{i=1}^{n} g(r_i) = (-1)^{mn} b_m^n \prod_{j=1}^{m} f(s_j).$$

Proof: In a splitting field of $f(x)g(x)$ over K, we have the factorizations

$$f(x) = a_n(x - r_1)(x - r_2)\ldots(x - r_n)$$
$$g(x) = b_m(x - s_1)(x - s_2)\ldots(x - s_m).$$

Thus $f(x)$ and $g(x)$ are obtained from

$$F(x) = a_n(x - u_1)(x - u_2)\ldots(x - u_n)$$
$$G(x) = b_m(x - y_1)(x - y_2)\ldots(x - y_m),$$

where $u_1, u_2, \ldots, u_n, y_1, y_2, \ldots, y_m$ are indeterminates over K, by substituting r_i for u_i and s_j for y_j. Since

$$R(F,G) = a_n^m b_m^n \prod_{i=1}^{n} \prod_{j=1}^{m} (u_i - y_j) = a_n^m \prod_{i=1}^{n} g(u_i) = (-1)^{mn} b_m^n \prod_{j=1}^{m} f(y_j)$$

by Theorem 56.5, this substitution gives

$$R(f,g) = a_n^m b_m^n \prod_{i=1}^{n} \prod_{j=1}^{m} (r_i - s_j) = a_n^m \prod_{i=1}^{n} g(r_i) = (-1)^{mn} b_m^n \prod_{j=1}^{m} f(s_j)$$ □

56.7 Lemma: *Let K be a field. Let*

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

be a polynomial of degree n in $K[x] \setminus K$ and

$$g(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0$$

a polynomial in $K[x] \setminus K$, possibly $b_m = 0$. Let $r_1, r_2, \ldots, r_n$ be the roots of $f(x)$ in some splitting field of $f(x)$ over K. Then

$$R(f,g) = a_n^m \prod_{i=1}^{n} g(r_i).$$

Proof: Assume first $b_m \neq 0$. Let F be a splitting field of $f(x)$ over K in which $r_1, r_2, \ldots, r_n$ lie and let E be a splitting field of $g(x)$ over F so that both $f(x)$ and $g(x)$ split completely in E. Then $R(f,g) = a_n^m \prod_{i=1}^{n} g(r_i)$ by Lemma 56.6.

Assume now $b_m = 0$ and let k be the largest index for which $b_k \neq 0$. Thus $b_m = b_{m-1} = \cdots = b_{k+1} = 0$ and $b_k \neq 0$. We put $G(x) = b_k x^k + b_{k-1}x^{k-1} + \cdots + b_1 x + b_0$. We get $R(f,g) = a_n^{m-k} R(f,G)$ from Remark 56.4 and we have $R(f,G) = a_n^k \prod_{i=1}^{n} G(r_i)$ by what we have just proved. Since $G(r_i) = g(r_i)$ for any $i = 1,2,\ldots,n$, we obtain

$$R(f,g) = a_n^{m-k} R(f,G) = a_n^{m-k}\, a_n^k \prod_{i=1}^{n} G(r_i) = a_n^m \prod_{i=1}^{n} G(r_i) = a_n^m \prod_{i=1}^{n} g(r_i).$$

This completes the proof. □

56.8 Definition: Let K be a field and $f(x)$ a nonzero polynomial in $K[x]$ of positive degree n. Let a_n be the leading coefficient of $f(x)$ and let $r_1, r_2, \ldots, r_n$ be the roots of $f(x)$ in some splitting field E of $f(x)$ over K. Then

$$a_n^{2n-2} \prod_{i<j} (r_i - r_j)^2 \in E$$

is called the *discriminant of* $f(x)$ and is denoted by $D(f)$.

It seems as though the discriminant of $f(x)$ depended on the splitting field E we choose and we had to call it actually the discriminant of $f(x)$ in E and denoted by $D_E(f)$. However, there is no need to refer to the splitting field since the discriminant is in fact an element of the field K. This we prove in the next theorem.

In the next theorem, if $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ and $f'(x) = n a_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \ldots + a_1$, then $R(f,f')$ is understood to be the determinant with $n + (n-1)$ rows, the first $n-1$ rows being

$$a_n\ a_{n-1} \cdots a_1\ a_0$$

surrounded with zeroes and the last n being

$$n a_n\ (n-1)a_{n-1} \cdots a_1$$

surrounded with zeroes, even if $n a_n = 0$, $(n-1)a_{n-1} = 0$, etc. (this happens when *char* $K = p \neq 0$ and $p \mid n$, $a_{n-1} = 0$, etc.). In other words, we define $R(f,f')$ as if f' is of degree $n-1$, although the degree of f' may be less than $n-1$ (cf. Remark 56.4).

56.9 Theorem: *Let K be a field and $f(x)$ a polynomial of positive degree n and let a_n be the leading coefficient of $f(x)$. Then the discriminant $D(f)$ of $f(x)$ is in K. In fact, $R(f,f') = (-1)^{n(n-1)/2} a_n D(f)$.*

Proof: Let E be a splitting field of $f(x)$ over K and let $r_1, r_2, \ldots, r_n$ be the roots of $f(x)$ in E. We evaluate $R(f,f')$. We have $R(f,f') = a_n^{n-1} \prod_{i=1}^{n} f'(r_i)$ by Lemma 56.7. We must find $f'(r_i)$. From $f(x) = a_n(x - r_1)(x - r_2)\ldots(x - r_n)$, we get

$$f'(x) = \sum_{j=1}^{n} a_n(x - r_1)\ldots(x - r_{j-1})(x - r_{j+1})\ldots(x - r_n)$$

so $$f'(r_i) = a_n(r_i - r_1)\ldots(r_i - r_{i-1})(r_i - r_{i+1})\ldots(r_i - r_n) = a_n \prod_{\substack{j=1 \\ j \neq i}}^{n} (r_i - r_j).$$

Thus $$R(f,f') = a_n^{n-1} \prod_{i=1}^{n} f'(r_i) = a_n^{n-1} \prod_{i=1}^{n} \left(a_n \prod_{\substack{j=1 \\ j \neq i}}^{n} (r_i - r_j)\right) = a_n^{2n-1} \prod_{i \neq j}^{n} (r_i - r_j)$$

$$= a_n \cdot a_n^{2n-2} \prod_{i \neq j} (r_i - r_j) = a_n \cdot a_n^{2n-2} \prod_{i<j} (r_i - r_j) \prod_{j<i} (r_i - r_j)$$

$$= a_n \cdot a_n^{2n-2} \prod_{i<j} (r_i - r_j) \prod_{j<i} (-1)(r_j - r_i)$$

$$= a_n \cdot a_n^{2n-2} \prod_{i<j} (r_i - r_j) \prod_{i<j} (-1)(r_i - r_j)$$

$$= a_n \cdot a_n^{2n-2} \prod_{i<j} (r_i - r_j) \cdot (-1)^{(n-1)+(n-2)+\cdots+2+1} \prod_{i<j} (r_i - r_j)$$

$$= (-1)^{n(n-1)/2} a_n \cdot a_n^{2n-2} \prod_{i<j} (r_i - r_j)^2 = (-1)^{n(n-1)/2} a_n D(f). \qquad \square$$

56.10 Examples: (a) Let K be a field and $ax^2 + bx + c \in K[x]$, with $a \neq 0$. The discriminant of $f(x)$ is $(-1)a^{-1}$ times the resultant

$$\begin{vmatrix} a & b & c \\ 2a & b & 0 \\ 0 & 2a & b \end{vmatrix} = \begin{vmatrix} a & b & c \\ 0 & -b & -2c \\ 0 & 2a & b \end{vmatrix} = a \begin{vmatrix} -b & -2c \\ 2 & b \end{vmatrix} = a(-b^2 + 4ac) = -a(b^2 - 4ac),$$

hence the discriminant of $f(x)$ is $b^2 - 4ac$.

(b) Let K be a field and $x^3 + px + q \in K[x]$. The discriminant of $f(x)$ is $(-1)^{3\cdot 2/2} 1^{-1}$ times the resultant

$$\begin{vmatrix} 1 & 0 & p & q & 0 \\ 0 & 1 & 0 & p & q \\ 3 & 0 & p & 0 & 0 \\ 0 & 3 & 0 & p & 0 \\ 0 & 0 & 3 & 0 & p \end{vmatrix} = \begin{vmatrix} 1 & 0 & p & q & 0 \\ 0 & 1 & 0 & p & q \\ 0 & 0 & -2p & -3q & 0 \\ 0 & 3 & 0 & p & 0 \\ 0 & 0 & 3 & 0 & p \end{vmatrix} = \begin{vmatrix} 1 & 0 & p & q \\ 0 & -2p & -3q & 0 \\ 3 & 0 & p & 0 \\ 0 & 3 & 0 & p \end{vmatrix}$$

$$= \begin{vmatrix} 1 & 0 & p & q \\ 0 & -2p & -3q & 0 \\ 0 & 0 & -2p & -3q \\ 0 & 3 & 0 & p \end{vmatrix} = \begin{vmatrix} 2p & 3q & 0 \\ 0 & 2p & 3q \\ 3 & 0 & p \end{vmatrix} = 4p^3 + 27q^2.$$

So the discriminant of $f(x)$ is equal to $-4p^3 - 27q^2$.

We now turn our attention to polynomial equations.

56.11 Lemma: (1) *Let $E/K, E_1/K_1$ be field extensions. Assume that there are field isomorphisms $\varphi: K \to K_1$ and $\psi: E \to E_1$ and that ψ is an extension of φ. Then $Aut_K E \cong Aut_{K_1} E_1$.*

(2) *Let K be a field and $f(x)$ a polynomial in $K[x]\backslash K$. Let E and F be two splitting fields of $f(x)$ over K. Then $Aut_K E \cong Aut_K F$.*

Proof: (1) For any $\sigma \in Aut_K E$, consider the mapping $\psi^{-1}\sigma\psi : E_1 \to E_1$. Clearly $\psi^{-1}\sigma\psi$ is a field isomorphism (Lemma 48.10). Moreover, for any $a_1 \in K_1$,

there is a unique $a \in K$ with $a\psi = a\varphi = a_1$, i.e., $a_1\varphi^{-1} = a_1\psi^{-1} = a$ and $a_1\psi^{-1}\sigma\psi = (a_1\psi^{-1})\sigma\psi = (a)\sigma\psi = (a\sigma)\psi = a\psi = a_1$, so $\psi^{-1}\sigma\psi$ is in fact a K_1-automorphism of E_1. Thus we have a mapping

$$\begin{aligned} A: Aut_K E &\to Aut_{K_1} E_1 \\ \sigma &\to \psi^{-1}\sigma\psi \end{aligned}$$

Now $(\sigma\tau)A = \psi^{-1}(\sigma\tau)\psi = (\psi^{-1}\sigma\psi)(\psi^{-1}\tau\psi) = \sigma A\, \tau A$ for any $\sigma,\tau \in Aut_K E$, so A is a group homomorphism. Repeating the same argument with K,E,φ,ψ and K_1, $E_1,\varphi^{-1},\psi^{-1}$ interchanged, we conclude that the mapping

$$\begin{aligned} B: Aut_{K_1} E_1 &\to Aut_K E \\ \theta &\to \psi\theta\psi^{-1} \end{aligned}$$

is an inverse of A, so A is one-to-one and onto $Aut_{K_1} E_1$. Thus A is an isomorphism and we get $Aut_K E \cong Aut_{K_1} E_1$.

(2) The fields E and F are K-isomorphic by Theorem 53.8, so the claim follows immediately from part (1). □

Thus Galois groups of any two splitting fields (over K) of $f(x)$ are isomorphic. This justifies the definite article in the next definition.

56.12 Definition: Let K be a field and $f(x)$ a polynomial in $K[x]\backslash K$. The Galois group $Aut_K E$ of a splitting field E of $f(x)$ over K is called the *Galois group of* $f(x) \in K[x]$.

56.13 Examples: (a) $\mathbb{Q}(i)$ is a splitting field of $x^2 + 1 \in \mathbb{Q}[x]$ over $\mathbb{Q}$ and hence the Galois group of $x^2 + 1 \in \mathbb{Q}[x]$ is $Aut_{\mathbb{Q}}\mathbb{Q}(i) \cong C_2$.

(b) The Galois group of $x^3 - 2$ is $\{\varphi_1,\varphi_2,\varphi_3,\varphi_4,\varphi_5,\varphi_6\} \cong S_3$. Here we used the notation of Example 54.18(a).

(c) The Galois group of $x^4 - 2$ is $\{\varphi_1,\varphi_2,\varphi_3,\varphi_4,\varphi_5,\varphi_6,\varphi_7,\varphi_8\} = \langle\sigma,\tau\rangle \cong D_8$. Here we used the notation of Example 54.18(b). We know that $D_8 \cong \{\iota,(13),(24),(12)(34),(13)(24),(14)(23),(1234),(1432)\} \leqslant S_4$.

(d) Let p be a prime number. The field $\mathbb{F}_{p^n}$ is a splitting field of $x^{p^n} - x$ over $\mathbb{F}_p$(Example 53.5(f)). Hence the Galois group of $x^{p^n} - x \in \mathbb{F}_p[x]$ is $Aut_{\mathbb{F}}\mathbb{F}_{p^n}[x] = \langle\sigma\rangle$, where σ is the homomorphism $a \to a^p$ (Example 54.18(c)).

56.14 Theorem: *Let K be a field, $f(x)$ a polynomial in $K[x]\backslash K$ and let G be the Galois group of $f(x)$. Then G is isomorphic to a subgroup of a symmetric group S_n.*

Proof: Let E be a splitting field of $f(x)$ over K and let $a_1, a_2, \ldots, a_n$ be the distinct roots of $f(x)$ in E ($1 \leqslant n \leqslant deg\ f(x)$). Any $\varphi \in G = Aut_K E$ maps any a_i to a a_j and thus gives rise to a permutation $\sigma_\varphi \in S_n$, namely $i \to j$. Thus σ_φ is given by $a_i\varphi \to a_{i\sigma_\varphi}$.

Now the mapping $\sigma: G \to S_n$ is a homomorphism of groups since, for any $\varphi \to \sigma_\varphi$

$\varphi, \psi \in G$, we have

$$a_{i\sigma_{\varphi\psi}} = a_i(\varphi\psi) = (a_i\varphi)\psi = a_{i\sigma_\varphi}\psi = a_j\psi \quad (\text{put } i\sigma_\varphi = j)$$
$$= a_{j\sigma_\psi} = a_{(i\sigma_\varphi)\sigma_\psi} = a_{i(\sigma_\varphi\sigma_\psi)}$$

for $i = 1,2,\ldots,n$ and so $\sigma_{\varphi\psi} = \sigma_\varphi\sigma_\psi$. Here $\varphi \in Ker\ \sigma$ if and only if $a_i\varphi = a_i$ for all $i = 1,2,\ldots,n$. Thus an automorphism in $Ker\ \sigma$ fixes each element of K and fixes each a_i. Since E is generated by a_i over K (Example 53.5(d)), we deduce that an automorphism in $Ker\ \sigma$ fixes all elements of E. Thus $Ker\ \sigma = \{\iota_E\}$. So σ is one-to-one and G is isomorphic to $Im\ \sigma \leqslant S_n$. □

The preceding proof is quite simple. G acts on the set of distinct roots of $f(x)$, and the permutation representation σ is one-to-one; thus G is isomorphic to a subgroup of S_U, and S_U itself is isomorphic to S_n. We will often identify the Galois group of a polynomial with its isomorphic images in S_U and in S_n.

The Galois group of a polynomial reflects many important properties of that polynomial. We describe how irreducibility is reflected in the Galois group. It turns out that the decomposition of $f(x)$ into irreducible polynomials is intimately connected with the partitioning of its roots into disjoint orbits. Let us recall that a group G is said to act transitively on a set X provided, for any $x,y \in X$, there is a $g \in G$ such that $xg = y$

(Definition 25.11). If $G \leqslant S_n$ acts transitively on $\{1,2,\ldots,n\}$, then we shall call G a *transitive subgroup of* S_n. Thus $G \leqslant S_n$ is transitive if and only if, for any $i,j \in \{1,2,\ldots,n\}$, there is a $\tau \in G$ such that $i\tau = j$.

56.15 Examples: (a) A subgroup G of S_n is transitive if and only if, for any $i \in \{1,2,\ldots,n\}$, there is a $\sigma \in S_n$ such that $1\sigma = i$. The necessity of this condition is clear. Conversely, if the condition is satisfied and i,j are in $\{1,2,\ldots,n\}$, there are $\sigma,\tau \in G$ with $1\sigma = i$ and $1\tau = j$, so $\sigma^{-1}\tau \in G$ maps i to j; hence the condition is also sufficient.

(b) If $H \leqslant G \leqslant S_n$ and H is transitive, then G is also transitive.

(c) $A_3 = \{\iota,(123),(132)\}$ is a transitive subgroup of S_3 for there are permutations σ_i in A_3 with $1\sigma_i = i$ for any $i = 1,2,3$, viz. $\sigma_1 = \iota$, $\sigma_2 = (123)$ and $\sigma_3 = (132)$. Then S_3 is of course another transitive subgroup of S_3. On the other hand, $\{\iota,(12)\}$ is not a transitive subgroup of S_3 for there is no permutation σ in $\{\iota,(12)\}$ that maps 1 to 3. Likewise $\{\iota,(13)\}$ and $\{\iota,(23)\}$ are not transitive subgroups of S_3. Certainly $\{\iota\}$ is not a transitive subgroup of S_3. Thus A_3 and S_3 are the only transitive subgroups of S_3.

(d) Let $\sigma = (12\ldots n) \in S_n$. Then $\langle\sigma\rangle$ is a transitive subgroup of S_n since $1\sigma^i = i$ for any $i = 1,2,\ldots,n$.

(e) If G is a transitive subgroup of S_n, so is any conjugate of G. Indeed, if G is transitive and $\tau \in S_n$, then, for any $i,j \in \{1,2,\ldots,n\}$, there is a $\sigma \in G$ that maps $i\tau^{-1}$ to $j\tau^{-1}$, i.e., $i\tau^{-1}\sigma\tau = j$. Thus there is a $\sigma^\tau \in G^\tau$ that maps i to j and G^τ is therefore transitive.

(f) It follows from the last two examples that $\langle(1234)\rangle$ and its conjugates $\langle(1324)\rangle,\langle(1243)\rangle$ are transitive subgroups of S_4. Also $V_4 = \{\iota,(12)(34),(13)(24),(14)(23)\}$ is a transitive subgroup of S_4. From $V_4 \leqslant A_4$ and $V_4 \leqslant S_4$, we see that A_4 and S_4 are transitive subgroups of S_4. Likewise $D = \{\iota,(13),(24),(12)(34),(13)(24),(14)(23),(1234),(1432)\}$ and its conjugates

$$\{\iota,(12),(34),(13)(24),(12)(34),(14)(32),(1324),(1423)\}$$
$$\{\iota,(14),(23),(12)(43),(14)(23),(13)(24),(1243),(1342)\}$$

are transitive subgroups of S_4. On the other hand, $\{\iota,(12),(34),(12)(34)\}$ and its conjugates are not transitive subgroups of S_4.

56.16 Theorem: *Let K be a field and let $f(x) \in K[x]$ be a monic polynomial having no multiple roots. Let E be a splitting field of $f(x)$ and $G = Aut_K E$ the Galois group of $f(x)$. Let $r_1, r_2, \ldots, r_n \in E$ be the roots of $f(x)$. Let $m_0 = 0$ and $m_k = n$.*

(1) *Assume the notation so chosen that*

$$\{r_1, r_2, \ldots, r_{m_1}\}, \{r_{m_1+1}, r_{m_1+2}, \ldots, r_{m_2}\}, \{r_{m_2+1}, r_{m_2+2}, \ldots, r_{m_3}\},$$
$$\ldots, \{r_{m_{k-1}+1}, r_{m_{k-1}+2}, \ldots, r_{m_k}\}$$

are the disjoint orbits under the action of G. Put

$$f_i(x) = (x - r_{m_{i-1}+1})(x - r_{m_{i-1}+2}) \ldots (x - r_{m_i}) \in E[x] \quad \text{for } i = 1,2, \ldots, k.$$

Then $f_i(x) \in K[x]$ and $f_i(x)$ is irreducible in $K[x]$, so that

$$f(x) = f_1(x) f_2(x) \ldots f_k(x)$$

is the canonical decomposition of $f(x)$ into irreducible polynomials in $K[x]$.

(2) *Let $f(x) = f_1(x) f_2(x) \ldots f_k(x)$ be the canonical decomposition of $f(x)$ into monic irreducible polynomials in $K[x]$ and let $r_{m_{i-1}+1}, r_{m_{i-1}+2}, \ldots, r_{m_i}$ be the roots of $f_i(x)$ $(i = 1,2, \ldots, k)$. Then*

$$\{r_1, r_2, \ldots, r_n\} = \{r_1, r_2, \ldots, r_{m_1}\} \cup \{r_{m_1+1}, r_{m_1+2}, \ldots, r_{m_2}\} \cup \{r_{m_2+1}, r_{m_2+2}, \ldots, r_{m_3}\}$$
$$\cup \ldots \cup \{r_{m_{k-1}+1}, r_{m_{k-1}+2}, \ldots, r_{m_k}\}$$

is the partitioning of $\{r_1, r_2, \ldots, r_n\}$ into disjoint orbits under the action of G.

Proof: (1) We first prove that $f_i(x) \in K[x]$. The coefficients of $f_i(x) = (x - r_{m_{i-1}+1})(x - r_{m_{i-1}+2}) \ldots (x - r_{m_i})$ are elementary symmetric polynomials in $r_{m_{i-1}+1}, r_{m_{i-1}+2}, \ldots, r_{m_i}$. Any automorphism in G maps each one of these $r_{m_{i-1}+1}, r_{m_{i-1}+2}, \ldots, r_{m_i}$ to one of them again and thus leaves the coefficients of $f_i(x)$ unchanged. So the coefficients of $f_i(x)$ are in the fixed field of G. Now $f(x)$ has no multiple roots, so the irreducible divisors of $f(x)$ are separable over K and, since E is a splitting field of $f(x)$ over K, we infer E is a Galois extension of K (Theorem 55.7) and the fixed field of G is exactly K. Hence $f_i(x) \in K[x]$.

We prove next that $f_i(x)$ is irreducible in $K[x]$. Let $g(x) \in K[x]$ be an irreducible divisor of $f_i(x)$. In E, there is a root of $g(x)$, say $r_{m_{i-1}+1}$. Then, for any $\varphi \in G$, $r_{m_{i-1}+1}\varphi$ is also a root of $g(x)$. But $\{r_{m_{i-1}+1}\varphi : \varphi \in G\}$ = orbit of $r_{m_{i-1}+1} = \{r_{m_{i-1}+1}, r_{m_{i-1}+2}, \ldots, r_{m_i}\}$. Thus each of $r_{m_{i-1}+1}, r_{m_{i-1}+2}, \ldots, r_{m_i}$ is a root of $g(x)$. These roots are distinct, for $f(x)$ has no multiple roots. Thus $g(x)$ has

at least $m_i - m_{i-1}$ distinct roots. Then $m_i - m_{i-1} \leq deg\ g(x) \leq deg\ f_i(x) = m_i - m_{i-1}$ and so $g(x) = f_i(x)$. Thus $f_i(x) = g(x)$ is irreducible in $K[x]$.

It follows that $f(x) = f_1(x)f_2(x)\ldots f_k(x)$ is the canonical decomposition of $f(x)$ into irreducible polynomials in $K[x]$.

(2) Suppose now $f(x) = f_1(x)f_2(x)\ldots f_k(x)$ is the canonical decomposition of $f(x)$ into irreducible polynomials in $K[x]$. We are to show that the roots of $f_i(x)$ make up the orbit of $r_{m_{i-1}+1}$. Indeed, if $\varphi \in G$, then $r_{m_{i-1}+1}\varphi$ is also a root of $f_i(x)$ and thus:

$$\text{orbit of } r_{m_{i-1}+1} \subseteq \{r_{m_{i-1}+1}, r_{m_{i-1}+2}, \ldots, r_{m_i}\}.$$

On the other hand, if $r \in E$ is any root of $f_i(x)$, then $K(r_{m_{i-1}+1}) \cong K(r)$ by a K-isomorphism ψ that sends $r_{m_{i-1}+1}$ to r (Theorem 53.2) and ψ can be extended to a K-automorphism φ (Theorem 53.7; E is a splitting field of $f(x)$ over $K(r_{m_{i-1}+1})$ and over $K(r)$ by Example 53.5(e)). So there is a $\varphi \in G$ with $r_{m_{i-1}+1}\varphi = r$ and any root r of $f_i(x)$ is in the orbit of $r_{m_{i-1}+1}$. Thus:

$$\{r_{m_{i-1}+1}, r_{m_{i-1}+2}, \ldots, r_{m_i}\} \subseteq \text{orbit of } r_{m_{i-1}+1}.$$

This completes the proof. □

56.17 Theorem: *Let K be a field, $f(x)$ a polynomial of positive degree n in $K[x]$ and let G be the Galois group of $f(x)$. If $f(x)$ is irreducible and separable over K, then n divides $|G|$ and G is isomorphic to a transitive subgroup of S_n.*

Proof: Let E be a splitting field of $f(x)$ over K. Then E is a Galois extension of K (Theorem 55.7) and, under the action of G, there is only one orbit of the roots of $f(x)$. Thus G acts transitively on the set of roots of $f(x)$ and its isomorphic image in S_n acts transitively on $\{1,2, \ldots,n\}$. So G is isomorphic to a transitive subgroup of S_n. Furthermore, if $r \in E$ is any root of $f(x)$, then $K(r)$ is an intermediate field of E/K and $|K(r):K| = deg\ f = n$ (Theorem 50.7) and, by the fundamental Theorem of Galois theory, G has a subgroup $K(r)'$ of index $|G:K(r)'| = |K(r):K| = n$. So n divides $|G|$ by Lagrange's theorem. □

We shall regard the Galois group as a subgroup of S_n. It will be interesting to determine the role of A_n. This is connected with discriminants.

56.18 Theorem: *Let K be a field such that char $K \neq 2$ and let $f(x) \in K[x]$. Assume deg $f = n > 0$ and let E be a splitting field of $f(x)$ over K. Suppose $f(x)$ has n distinct roots $r_1, r_2, \ldots, r_n$ in E. Put*

$$\delta = \prod_{i<j}(r_i - r_j) = (r_1 - r_2)(r_1 - r_3)\ldots(r_{n-1} - r_n) \text{ and } d = \delta^2.$$

(1) *For $\varphi \in Aut_K E \leqslant S_n$, there holds $\delta\varphi = \delta$ if and only if φ is in A_n and $\delta\varphi = -\delta$ if and only if φ is in $S_n \backslash A_n$.*

(2) *d, which is an elements of E, is actually in K. In fact, $d = a_n^{-(2n-2)}D(f)$, where a_n is the leading coefficient and $D(f)$ is the discriminant of $f(x)$.*

Proof: (1) We have $\delta\varphi = \prod_{i<j}(r_{i'} - r_{j'}) = (r_{1'} - r_{2'})(r_{1'} - r_{3'})\ldots(r_{(n-1)'} - r_{n'})$, where $r_{i'} = r_i\varphi$. We divide the ordered pairs (i,j) with $i < j$ into two classes according as $i' < j'$ or $i' > j'$. Then $\delta\varphi = \prod_{\substack{i<j \\ i'<j'}}(r_{i'} - r_{j'}) \prod_{\substack{i<j \\ i'>j'}}(r_{i'} - r_{j'})$

$$= \prod_{\substack{i<j \\ i'<j'}}(r_{i'} - r_{j'}) \prod_{\substack{i<j \\ i'>j'}}(-1)(r_{j'} - r_{i'})$$

$$= \prod_{\substack{i<j \\ i'<j'}}(r_{i'} - r_{j'}) \prod_{\substack{j<i \\ j'>i'}}(-1)(r_{i'} - r_{j'}) \qquad \text{(interchange the dummy indices } i \text{ and } j\text{)}$$

$$= \prod_{\substack{i<j \\ i'<j'}}(r_{i'} - r_{j'}) \cdot (-1)^s \prod_{\substack{j<i \\ j'>i'}}(r_{i'} - r_{j'}) \qquad \text{(where } s \text{ is the number of factors in the second product; hence } s \text{ is the number of inversions of the permutation } \begin{pmatrix} 1 & 2 & \ldots & n \\ 1' & 2' & \ldots & n' \end{pmatrix} = \varphi \in Aut_K E \leqslant S_n\text{)}$$

$$= (-1)^s \prod_{\substack{i<j \\ i'<j'}}(r_{i'} - r_{j'}) \prod_{\substack{j<i \\ j'>i'}}(r_{i'} - r_{j'}) = \varepsilon(\varphi) \prod_{i'<j'}(r_{i'} - r_{j'}) = \varepsilon(\varphi) \prod_{i<j}(r_i - r_j) = \varepsilon(\varphi)\delta.$$

This proves (1).

(2) The equation $d = a_n^{-(2n-2)}D(f)$ is immediate from the definition of discriminant (Definition 56.8). This implies of course that d is in K, since $D(f)$, being a_n^{-1} times a determinant of a matrix with entries in K, is an element of K. Alternatively, we have $\delta\varphi = \mp\delta$ and thus $d\varphi = (\delta^2)\varphi = (\delta\varphi)^2 = (\mp\delta)^2 = \delta^2 = d$ for any $\varphi \in Aut_K E$. So d is in the fixed field of $Aut_K E$. Since the roots of $f(x)$ are simple by hypothesis, the irreducible divisors of $f(x)$ are separable over K and thus E is Galois over K (Theorem 55.7), so the fixed field of $Aut_K E$ is K and d is in K. □

56.19 Theorem: *Let K be a field such that char $K \neq 2$ and let $f(x) \in K[x]$. Assume deg $f = n > 0$ and let E be a splitting field of $f(x)$ over K. Suppose $f(x)$ has n distinct roots $r_1, r_2, \ldots, r_n$ in E so that E is a Galois extension of K* (Theorem 55.7). *Put $\delta = \prod_{i<j}(r_i - r_j)$. Consider the Galois group $Aut_K E$ as a subgroup of S_n.*

In the Galois correspondence, the intermediate field $K(\delta)$ corresponds to $Aut_K E \cap A_n$. In particular, $Aut_K E \leqslant A_n$ if and only if $\delta \in K$.

Proof: In the Galois correspondence, the subgroup of $Aut_K E$ corresponding to the intermediate field $K(\delta)$ is

$$\begin{aligned} K(\delta)' &= \{\varphi \in Aut_K E: a\varphi = a \text{ for all } a \in K(\delta)\} \\ &= \{\varphi \in Aut_K E: \delta\varphi = \delta\} \\ &= \{\varphi \in Aut_K E: \varphi \in A_n\} \\ &= Aut_K E \cap A_n \end{aligned}$$

by Theorem 56.18. In particular, $Aut_K E \leqslant A_n$ if and only if $Aut_K E \cap A_n = Aut_K E$, so if and only if $K(\delta)' = Aut_K E = K'$, hence if and only if $K(\delta) = K$, hence if and only if $\delta \in K$. □

We now study Galois groups of polynomials of degree 2,3,4. We start with quadratic polynomials.

56.20 Theorem: *Let K be a field and $f(x)$ an irreducible polynomial in $K[x]$ of degree 2. Let G be the Galois group of $f(x)$, regarded as a subgroup of S_2. If $f(x)$ is separable over K, then $G = S_2 \cong C_2$. If $f(x)$ is not separable over K, then $G = 1$.*

Proof: If $f(x)$ is separable over K, then G is a transitive subgroup of S_2 (Theorem 56.17). Since S_2 is the only transitive subgroup of S_2, the result follows. If $f(x) = ax^2 + bx + c$ is not separable over K, then $f'(x) = 2ax + b = 0$, so $2a = 0 = b$ (and $a \neq 0$), so *char* $K = 2$ and $f(x) = a(x^2 + e)$ for some suitable $e \in K$, and a splitting field of $f(x)$ over K is $K(r)$, where r is a root of $f(x)$. Then any φ in G maps r to r and thus fixes $K(r)$. This means G consists of the identity mapping on $K(r)$. Hence $G = 1$. □

56.21 Theorem: *Let K be a field and $f(x)$ an irreducible separable polynomial in $K[x]$ of degree 3. Let G be the Galois group of $f(x)$, regarded as a subgroup of S_3. Then $G = S_3$ or $G = A_3$. More specifically, if char $K \neq 2$, then $G = A_3$ in case $D(f)$ is the square of an element in K, and $G = S_3$ in case $D(f)$ is not the square of any element in K.*

Proof: G is a transitive subgroup of S_3 (Theorem 56.17). Since S_3 and A_3 are the only transitive subgroups of S_3 (Example 56.15(c)), the result follows.

Assume in addition *char* $K \neq 2$: Then $G = A_3$ if and only if $\delta \in K$ in the notation of Theorem 56.19. Since $\delta^2 = a_3^{-4}D(f)$, where a_3 is the leading coefficient of $f(x)$ (Theorem 56.18), we conclude $G = A_3$ if and only if $a_3^{-4}D(f)$ is the square of an element in K, thus if and only if $D(f)$ is the square of an element in K. □

56.22 Examples: (a) Let $x^3 + 6x + 2 \in \mathbb{F}_7[x]$. This polynomial has no root in $\mathbb{F}_7$, hence is irreducible and then clearly separable over $\mathbb{F}_7$. Its discriminant $-4(6)^3 - 27(2)^2 = 4 + 1 \cdot 4 = 1 = 1^2$ (Example 56.10(b)) is a square in $\mathbb{F}_7$, so the Galois group of $x^3 + 6x + 2$ is A_3.

(b) Let $x^3 + 5x + 5 \in \mathbb{Q}[x]$. This polynomial is irreducible by Eisenstein's criterion and is separable over $\mathbb{Q}$ since *char* $\mathbb{Q} = 0$. The discriminant is

equal to $-4(5)^3 - 27(5)^2 = -1175$, which is not a square in $\mathbb{Q}$. So the Galois group of $x^3 - 5x + 5$ is S_3.

Next we investigate polynomials of degree four. Here S_4 will come into play. We know that $V_4 = \{1, (12)(34), (13)(24), (14)(23)\}$ is an important normal subgroup of S_4. It will be useful to find the intermediate field corresponding to V_4 in the Galois correspondence.

56.23 Theorem: *Let K be a field such that char $K \neq 2$ and let $f(x) \in K[x]$ be a polynomial of degree four. Let E be a splitting field of $f(x)$ over K. Suppose $f(x)$ has four distinct roots r_1, r_2, r_3, r_4 in E so that E is a Galois extension of K* (Theorem 55.7). *We put $\alpha = r_1r_2 + r_3r_4$, $\beta = r_1r_3 + r_2r_4$ and $\gamma = r_1r_4 + r_2r_3$ and consider the Galois group $Aut_K E$ as a subgroup of S_4* (Theorem 56.14).
In the Galois correspondence, the intermediate field $K(\alpha,\beta,\gamma)$ corresponds to $Aut_K E \cap V_4$.

Proof: In the Galois correspondence, the subgroup of $Aut_K E$ corresponding to the intermediate field $K(\alpha,\beta,\gamma)$ is

$$K(\alpha,\beta,\gamma)' = \{\varphi \in Aut_K E: a\varphi = a \text{ for all } a \in K(\alpha,\beta,\gamma)\}$$
$$= \{\varphi \in Aut_K E: \alpha\varphi = \alpha, \beta\varphi = \beta, \gamma\varphi = \gamma\}.$$

If $\varphi = (12)(34) \in Aut_K E$, then φ fixes α since $\alpha\varphi = (r_1r_2 + r_3r_4)\varphi = r_2r_1 + r_4r_3 = r_1r_2 + r_3r_4 = \alpha$. Similarly $\beta\varphi = (r_1r_3 + r_2r_4)\varphi = r_2r_4 + r_1r_3 = \beta$ and $\gamma\varphi = (r_1r_4 + r_2r_3)\varphi = r_2r_3 + r_1r_4 = \gamma$. Thus $(12)(34) \in K(\alpha,\beta,\gamma)'$ if $(12)(34)$ is in $Aut_K E$. In like manner, one verifies that $(13)(24)$ and $(14)(23)$ belong to $K(\alpha,\beta,\gamma)'$ whenever they are in $Aut_K E$. This proves $V_4 \cap Aut_K E \leq K(\alpha,\beta,\gamma)'$.

To complete the proof, we show, for any $\varphi \in Aut_K E$, that $\varphi \notin V_4$ implies $\varphi \notin K(\alpha,\beta,\gamma)'$. Indeed if $\varphi \notin V_4$, then φ is in one of the cosets $V_4(12)$, $V_4(13)$, $V_4(23)$, $V_4(123)$, $V_4(132)$ of V_4 in S_4. If $\varphi \in V_4(12)$, then $\varphi = \psi(12)$ for some $\psi \in V_4 \cap Aut_K E$, therefore $(r_1r_3 + r_2r_4)\varphi = (r_1r_3 + r_2r_4)\psi(12) = (r_1r_3 + r_2r_4)(12)$ and φ does not fix β since

$$r_1r_3 + r_2r_4 = \beta = \beta\varphi = (r_1r_3 + r_2r_4)\varphi = (r_1r_3 + r_2r_4)(12) = r_2r_3 + r_1r_4$$

yields $(r_1 - r_2)r_3 = (r_1 - r_2)r_4$ and so $r_1 = r_2$ or $r_3 = r_4$, contrary to the hypothesis that the roots of $f(x)$ are distinct. Similarly, if $\varphi \in V_4(13)$, then

φ does not fix γ and if $\varphi \in V_4(23)$, then φ does not fix α. If $\varphi \in V_4(123)$, then φ does not fix α since

$$r_1r_2 + r_3r_4 = \alpha = \alpha\varphi = (r_1r_2 + r_3r_4)\varphi = (r_1r_2 + r_3r_4)(123) = r_2r_3 + r_1r_4$$

yields $(r_1 - r_3)r_2 = (r_1 - r_3)r_4$ and so $r_1 = r_3$ or $r_2 = r_4$, contrary to the hypothesis. Similarly, if $\varphi \in V_4(132)$, then φ does not fix α. This proves that no automorphism in $Aut_K E \setminus V_4$ can be in $K(\alpha,\beta,\gamma)'$. Hence we obtain $K(\alpha,\beta,\gamma)' \leqslant V_4 \cap Aut_K E$, as was to be proved. □

56.24 Definition: Let K be a field and let $f(x) \in K[x]$ be a polynomial of degree four having four distinct roots r_1,r_2,r_3,r_4 in a splitting field of $f(x)$ over K. We put $\alpha = r_1r_2 + r_3r_4$, $\beta = r_1r_3 + r_2r_4$ and $\gamma = r_1r_4 + r_2r_3$. The polynomial $(x - \alpha)(x - \beta)(x - \gamma) \in K(\alpha,\beta,\gamma)[x]$ is called the *resolvent cubic of* $f(x)$.

56.25 Lemma: *Let K be a field and let $f(x) \in K[x]$ be a polynomial of degree four having four distinct roots in a splitting field of $f(x)$ over K. Then the resolvent cubic of $f(x)$ is a polynomial in $K[x]$. In fact, if $f(x) = x^4 + bx^3 + cx^2 + dx + e$, then the resolvent cubic of $f(x)$ is equal to*

$$x^3 - cx^2 + (bd - 4e)x - (b^2e - 4ce + d^2).$$

Proof: This is routine computation. Let r_1,r_2,r_3,r_4 be the roots of $f(x)$ in a splitting field of $f(x)$ over K. The resolvent cubic of $f(x)$ is

$$x^3 - (\alpha + \beta + \gamma)x^2 + (\alpha\beta + \alpha\gamma + \beta\gamma)x - (\alpha\beta\gamma),$$

where $\alpha = r_1r_2 + r_3r_4$, $\beta = r_1r_3 + r_2r_4$, $\gamma = r_1r_4 + r_2r_3$. Let σ_m be the m-th elementary symmetric polynomial in 4 indeterminates. Then we have

$$\alpha + \beta + \gamma = r_1r_2 + r_3r_4 + r_1r_3 + r_2r_4 + r_1r_4 + r_2r_3 = \sigma_2(r_1,r_2,r_3,r_4) = c;$$

$$\alpha\beta + \alpha\gamma + \beta\gamma = r_1{}^2r_2r_3 + \cdots = \cdots$$
$$= (r_1 + r_2 + r_3 + r_4)(r_1r_2r_3 + r_1r_2r_4 + r_1r_3r_4 + r_2r_3r_4) - 4r_1r_2r_3r_4 =$$
$$= \sigma_1(r_1,r_2,r_3,r_4)\sigma_3(r_1,r_2,r_3,r_4) - \sigma_4(r_1,r_2,r_3,r_4) = bd - 4e;$$

$$\alpha\beta\gamma = \cdots = b^2e - 4ce + d^2.$$ □

56.26 Theorem: *Let K be a field and let $f(x) \in K[x]$ be a polynomial of degree four, which is irreducible and separable over K. Let E be a splitting field of $f(x)$ over K and let r_1,r_2,r_3,r_4 be the (distinct) roots of $f(x)$ in E. We put $\alpha = r_1r_2 + r_3r_4$, $\beta = r_1r_3 + r_2r_4$ and $\gamma = r_1r_4 + r_2r_3$. Let $G = Aut_K E$*

be the Galois group of $f(x)$, considered as a subgroup of S_4. We put $|K(\alpha,\beta,\gamma):K| = m$. Then G can be described as follows.

$$\begin{aligned} G = S_4 &\iff m = 6. \\ G = A_4 &\iff m = 3. \\ G \cong D_8 &\iff m = 2 \text{ and } f(x) \text{ is irreducible over } K(\alpha,\beta,\gamma). \\ G = V_4 &\iff m = 1. \\ G \cong C_4 &\iff m = 2 \text{ and } f(x) \text{ is reducible over } K(\alpha,\beta,\gamma). \end{aligned}$$

Proof: Since $f(x)$ is irreducible and separable over K, its roots are distinct. We know that G is a transitive subgroup of of S_4 and 4 divides $|G|$ (Theorem 56.17). The transitive subgroups of S_4 whose orders are divisible by 4 are S_4, A_4, the Sylow 2-subgroups of S_4 (isomorphic to D_8), V_4 and the cyclic groups generated by 4-cycles like $\langle(1234)\rangle$ (Example 56.15(f)). Thus G is one of S_4, A_4, D_8, V_4, C_4.

The intermediate field $K(\alpha,\beta,\gamma)$ corresponds to $V_4 \cap G$ (Theorem 56.23). Now E is Galois over $K(\alpha,\beta,\gamma)$ and the Galois group $Aut_{K(\alpha,\beta,\gamma)}E = K(\alpha,\beta,\gamma)'$ is $V_4 \cap G$. Since $V_4 \trianglelefteq S_4$, we have $V_4 \cap G \trianglelefteq G$ and so $K(\alpha,\beta,\gamma)$ is a Galois extension of K and the Galois group of $K(\alpha,\beta,\gamma)$ over K is (isomorphic to) $G/(G \cap V_4)$ (Theorem 54.25(2)). We get

$$m = |K(\alpha,\beta,\gamma):K| = |Aut_K K(\alpha,\beta,\gamma)| = |G/(G \cap V_4)| \text{ and}$$

$$\begin{aligned} G = S_4 &\implies m = |G/(G \cap V_4)| = |S_4/V_4| = 6; \\ G = A_4 &\implies m = |G/(G \cap V_4)| = |A_4/V_4| = 3; \\ G \cong D_8 &\implies m = |G/(G \cap V_4)| = |D_8/V_4| = 2; \text{ moreover, } E \text{ is a} \end{aligned}$$

splitting field of $f(x)$ over $K(\alpha,\beta,\gamma)$ and $Aut_{K(\alpha,\beta,\gamma)}E = K(\alpha,\beta,\gamma)' = V_4 \cap D_8 = V_4$ is a transitive subgroup of S_4, so $f(x)$ is irreducible over $K(\alpha,\beta,\gamma)$ by Theorem 56.16;

$$\begin{aligned} G = V_4 &\implies m = |G/(G \cap V_4)| = |V_4/V_4| = 1; \\ G \cong C_4 &\implies m = |G/(G \cap V_4)| = \\ &\qquad = |\{\iota,(1234),(13)(24),(1432)\}/\{\iota,(13)(24)\}| = 2 \end{aligned}$$

(eventually after renaming the roots, we may assume, without loss of generality, that $G = \{\iota,(1234),(13)(24),(1432)\}$); moreover, $Aut_{K(\alpha,\beta,\gamma)}E = K(\alpha,\beta,\gamma)' = \langle(1234)\rangle \cap V_4 = \langle(13)(24)\rangle$ is not a transitive subgroup of S_4, so $f(x)$ is not irreducible over $K(\alpha,\beta,\gamma)$ by Theorem 56.16.

This proves the $\implies$ assertions in the statement of the theorem. As the five cases are mutually exclusive, the converse assertions are also valid. □

56.27 Examples: (a) The polynomial $f(x) = x^4 - 4x^2 + 1 \in \mathbb{Q}[x]$ has no integer roots and is easily verified to have no quadratic factors in $\mathbb{Z}[x]$, so $f(x)$ is irreducible over $\mathbb{Z}$ and over $\mathbb{Q}$ (Lemma 34.11). Since *char* $\mathbb{Q} = 0$, $f(x)$ is separable over $\mathbb{Q}$. In order to determine its Galois group G, we find the resolvent cubic of $f(x)$. The resolvent cubic of $f(x)$ is

$$\begin{aligned} &= x^3 - (-4)x^2 + (0\cdot 0 - 4\cdot 1)x - (0^2\cdot 1 - 4(-4)(1) + 0^2) \\ &= x^3 + 4x^2 - 4x - 16 \\ &= (x + 4)(x - 2)(x + 2) \end{aligned}$$

and the roots α,β,γ of the resolvent cubic are $-4, -2, 2$. Thus $\mathbb{Q}(\alpha,\beta,\gamma) = \mathbb{Q}$ and $m = |\mathbb{Q}(\alpha,\beta,\gamma):\mathbb{Q}| = 1$. Theorem 56.26 yields $G = V_4$.

From $f(r) = 0 \Leftrightarrow (r^2 - 2)^2 = 3$, we see that the roots (say in $\mathbb{R}$) of $f(x)$ are

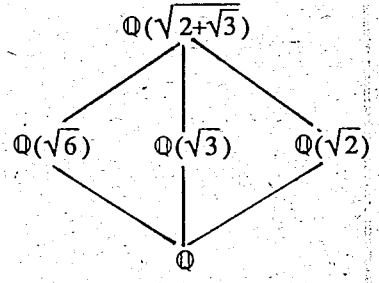
$$r_1 = \sqrt{2+\sqrt{3}},\quad r_2 = \sqrt{2-\sqrt{3}},\quad r_3 = -\sqrt{2+\sqrt{3}},\quad r_4 = -\sqrt{2-\sqrt{3}}.$$

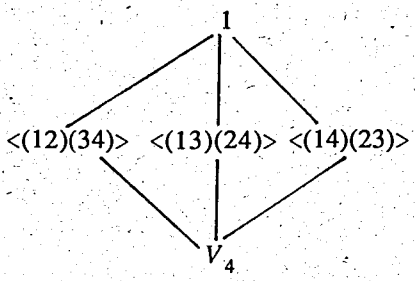
Note that $r_2 = 1/r_1$, $r_3 = -r_1$ and $r_4 = -1/r_1$. Since

$(12)(34) \in V_4 = G$ fixes $r_1 + r_2 = \sqrt{6}$,

$(13)(24) \in V_4 = G$ fixes r_1^2, hence also $r_1^2 - 2 = \sqrt{3}$,

$(14)(23) \in V_4 = G$ fixes $r_1 + r_4 = \sqrt{2}$, the Galois correspondence is as depicted below.





(b) Let $f(x) = x^4 + 5x^2 + 5 \in \mathbb{Q}[x]$. Then $f(x)$ is irreducible over $\mathbb{Z}$ by Eisenstein's criterion and also over $\mathbb{Q}$ by Lemma 34.11. Thus $f(x)$ is separable over $\mathbb{Q}$. Let G be the Galois group of $f(x)$. The resolvent cubic of $f(x)$ is $x^3 - 5x^2 - 20x + 100 = (x - 5)(x^2 - 20) = (x - 5)(x - 2\sqrt{5})(x + 2\sqrt{5})$, with roots $\alpha,\beta,\gamma = 5, 2\sqrt{5}, -2\sqrt{5}$. Hence $\mathbb{Q}(\alpha,\beta,\gamma) = \mathbb{Q}(\sqrt{5})$. So Theorem 56.26 gives $G \cong D_8$ or $G \cong C_4$. In fact, since

$$f(x) = \left(x^2 + \frac{5+\sqrt{5}}{2}\right)\left(x^2 + \frac{5-\sqrt{5}}{2}\right)$$

is reducible over $\mathbb{Q}(\sqrt{5})$, we have $G \cong C_4$.

(c) Let $f(x) = x^4 - 2 \in \mathbb{Q}[x]$. Then $f(x)$ is irreducible over $\mathbb{Q}$ by Eisenstein's criterion and Lemma 34.11. Let G be the Galois group of $f(x)$. The resolvent cubic of $f(x)$ is $x^3 + 8x$, whose roots are $\alpha,\beta,\gamma = 0, 2\sqrt{2}i, -2\sqrt{2}i$. Therefore $m = |\mathbb{Q}(\sqrt{2}i):\mathbb{Q}| = 2$ and $G \cong D_8$ or $G \cong C_4$. It is easy to see that $f(x)$ is irreducible over $\mathbb{Q}(\sqrt{2}i)$, so we get $G \cong D_8$ from Theorem 56.26.

Exercises

1. Find the resultant $R(f,g)$ when $f(x) = x^3 + 4x^3 - 3x^2 + x - 2 \in \mathbb{Q}[x]$ and $g(x) = x - 3 \in \mathbb{Q}[x]$.

2. Let K be a field and $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, $g(x) = b_1 x + b_0$ polynomials in $K[x]$, with $b_1 \neq 0$. Show that $R(f,g) = (-b_1)^n f(-b_1/b_0)$.

3. Let K be a field and $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, $g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$. If $n \geqslant m$, show that $R(f + cg,g) = R(f,g)$ for all $c \in K$.

4. Let K be a field and $f,g,h \in K[x]$. Prove that $R(fh,g) = R(f,g)R(h,g)$.

5. Let K be a field and $f,g,h \in K[x]$. Prove that $D(fg) = D(f)D(g)[R(f,g)]^2$ and that $D(f(x)) = D(f(x - c))$ for any $c \in K$.

6. Let K be a field and $f(x) = ax^3 + bx^2 + cx + d \in K[x]$. Prove that
$$D(f) = b^2c^2 + 18abcd - 4ac^3 - 4b^3d - 27a^2d^2.$$

7. Let K be a field and $f(x) = x^4 + ax^2 + bx + c \in K[x]$. Prove that
$$D(f) = -4a^3b^2 + 144acb^2 + 16a^4c - 128a^2c^2 + 256c^3 - 27b^4.$$

8. Let K be a field, $f(x)$ a polynomial of degree n in $K[x]$, with leading coefficient a_n and let $r_1,r_2,\ldots,r_n$ be the roots of $f(x)$ in some splitting field of $f(x)$ over K. Put $s_0 = n$ and $s_m = r_1^m + r_1^m + \cdots + r_n^m$ for $m \in \mathbb{N}$. Show that

$$D(f) = a^{2n-2}\begin{vmatrix} s_0 & s_1 & s_2 & \cdots & s_{n-1} \\ s_1 & s_2 & s_3 & \cdots & s_n \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ s_{n-1} & s_n & s_{n+1} & \cdots & s_{2n-2} \end{vmatrix}.$$

(Hint: multiply two Vandermonde determinants.)

8. Where did we use the hypothesis *char* $K \neq 2$ in Theorem 56.18?

9. Find the discriminants and Galois groups of the following polynomials.
   (a) $x^3 + 3x^2 - 1 \in \mathbb{Q}[x]$.
   (b) $x^3 - 2x^2 + 4x + 6 \in \mathbb{Q}[x]$.
   (c) $x^3 - x + 2 \in \mathbb{F}_3[x]$.
   (d) $x^3 + 3x^2 - 3 \in \mathbb{F}_5[x]$.

10. Find the Galois groups of the following polynomials over the fields indicated.
   (a) $x^4 - 2$ over $\mathbb{Q}(\sqrt{2})$ and over $\mathbb{Q}(\sqrt{2}i)$.
   (b) $(x^3 - 2)(x^2 - 5)$ over $\mathbb{Q}$.
   (c) $x^4 - 8x^2 + 15$ over $\mathbb{Q}$.
   (d) $x^4 + 4x^2 + 2$ over $\mathbb{Q}$ and over $\mathbb{Q}(\sqrt{2})$.
   (e) $(x^2 - 2)(x^2 - 3)(x^2 - 5)$ over $\mathbb{Q}$, over $\mathbb{Q}(\sqrt{2})$, over $\mathbb{Q}(\sqrt{6})$ and over $\mathbb{Q}(\sqrt{2},\sqrt{3})$.

11. Let K be any arbitrary field and $f(x) = x^3 - 3x + 1 \in K[x]$. Show that $f(x)$ is either irreducible over K or splits in K.

12. Let K be a field and $f(x)$ an irreducible separable polynomial of degree three in $K[x]$. Suppose r_1, r_2, r_3 are the roots of $f(x)$ in some splitting field of $f(x)$ over K. If the Galois group of $f(x)$ is S_3, show that, in the Galois correspondence, $K(r_i)$ corresponds to the subgroup $\{\iota, (jk)\}$ of S_3, where $\{i,j,k\} = \{1,2,3\}$.

13. Prove that S_4 has no transitive subgroup of order six.

14. Let p be a prime number and $G \leqslant S_p$. Show that G is transitive if and only if p divides the order of G.

§57
Norm and Trace

In this paragraph, we introduce norm and trace of elements in an extension field. These can be defined for any finite dimensional extension, but we restrict ourselves to the important case where the extension is separable.

In order to define norm and trace, we need K-homomorphisms of an extension field of K. In the case of a separable extension, these are easy to describe.

Let K be a field and E a finite dimensional separable extension of K and N a normal closure of K over E so that N is finite dimensional and Galois over K (Theorem 55.11). Let us put $|E:K| = n$. Since E is finite dimensional and hence finitely generated (Theorem 50.10) over K, there is an $a \in K$ such that $E = K(a)$ (Theorem 55.14). Let $f(x) \in K[x]$ be the (separable) minimal polynomial of a over K, so that $deg\ f(x) = |K(a):K| = |E:K| = n$ (Theorem 50.7). Since N is normal over K and $f(x)$ has a root a in N, the polynomial $f(x)$ splits in N, say

$$f(x) = (x - a_1)(x - a_2)\dots(x - a_n), \qquad a_1 = a,\ a_1, a_2, \dots, a_n \in N$$

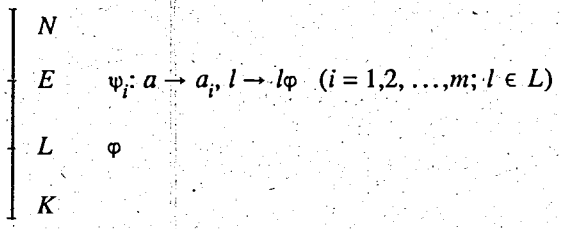
and $a_1, a_2, \dots, a_n$ are pairwise distinct. Then $K(a)$ and $K(a_i) \subseteq N$ are K-isomorphic by Theorem 53.2, namely by the K-isomorphism

$$\begin{array}{ccc} N & & N \\ \cup & & \cup \\ \psi_i:\ E = K(a) = K[a] & \longrightarrow & K[a_i] = K(a_i) \\ k_0 + k_1 a + \cdots + k_{n-2}a^{n-2} + k_{n-1}a^{n-1} & \longrightarrow & k_0 + k_1 a_i + \cdots + k_{n-2}a_i^{n-2} + k_{n-1}a_i^{n-1}, \end{array}$$

where $k_0, k_1, \dots, k_{n-2}, k_{n-1} \in K$. Thus each ψ_i is a K-homomorphisms from E into N. Conversely, any K-homomorphism $\psi: E \to N$ must map a to one of $a_1, a_2, \dots, a_n$ and must coincide with one of $\psi_1, \psi_2, \dots, \psi_n$. So $\{\psi_1, \psi_2, \dots, \psi_n\}$ is the complete set of K-homomorphisms from E into N.

We give a generalization of this result.

57.1 Lemma: *Let K be a field and E a finite dimensional separable extension of K. Let L be an intermediate field of E/K and let N be a normal closure of K over E. If $\varphi: L \to N$ is a K-homomorphism, then φ can be extended is exactly $|E:L|$ ways to a K-homomorphism $E \to N$.*



Proof: Since E is finitely generated and separable over K, it is a simple extension of K, say $E = K(a)$ (Theorem 55.14). Let $|E:L| = m$ and $g(x) \in L[x]$ the minimal polynomial of a over L so that $deg\ g(x) = m$. Let $f(x)$ be the minimal polynomial of a over K. Then $f(x)$ splits in N because the irreducible polynomial $f(x) \in K[x]$ has a root in N and N is normal over K. Since $g(x)$ divides $f(x)$ (in $L[x]$; Lemma 50.5), the roots of $g(x)$ are all in N. Let $a = a_1, a_2, \ldots, a_m \in N$ be the roots of $g(x)$. Then any extension $\psi: E \to N$ (ψ a K-homomorphism) of φ must send a to one of $a_1, a_2, \ldots, a_m$ and any l in L to $l\varphi$, and thus must coincide with one of the mappings

$$\psi_i: E = L(a) = L[a] \longrightarrow L[a_i] = L(a_i) \subseteq N$$
$$l_0 + l_1 a + \cdots + l_{m-2}a^{m-2} + l_{m-1}a^{m-1} \to (l_0\varphi) + (l_1\varphi)a_i + \cdots + (l_{m-2}\varphi)a_i^{m-2} + (l_{m-1}\varphi)a_i^{m-1}$$

($l_0, l_1, \ldots, l_{m-2}, l_{m-1} \in L$), where $i = 1,2,\ldots,m$; and these mappings ψ_i are indeed extensions of φ (since $\psi_i: l_0 \to l_0\varphi$) and field homomorphisms (cf. Lemma 53.1, Theorem 53.2). Thus $\{\psi_1, \psi_2, \ldots, \psi_m\}$ is the complete set of K-homomorphisms from E into N which are extensions of φ. □

We can now give the definition of norm and trace.

57.2 Definition: Let K be a field and E a finite dimensional separable extension of K. Let $a \in E$. Choose a normal closure N of K over E and let $\{\varphi_1, \varphi_2, \ldots, \varphi_k\}$ be the set of all K-homomorphisms from E into N (so $k = |E:K|$). The *norm of $a \in E$ over K* is denoted by $N_{E/K}(a)$ and is defined as

$$N_{E/K}(a) = (a\varphi_1)(a\varphi_2)\dots(a\varphi_k).$$

The *trace of* $a \in E$ *over* K is denoted by $T_{E/K}(a)$ and is defined as

$$T_{E/K}(a) = a\varphi_1 + a\varphi_2 + \cdots + a\varphi_k.$$

$N_{E/K}(a)$ and $N_{E/K}(a)$ therefore depend on E, K as well as a. It seems as though $N_{E/K}(a)$ and $T_{E/K}(a)$ depended also on the normal closure N we choose, but they actually do not depend on N. This will be proved shortly (Lemma 57.4(3)).

In case E is Galois over K, the normal closure N is equal to E and then we have

$$N_{E/K}(a) = (a\varphi_1)(a\varphi_2)\dots(a\varphi_k), \quad T_{E/K}(a) = a\varphi_1 + a\varphi_2 + \cdots + a\varphi_k,$$

where $\{\varphi_1, \varphi_2, \dots, \varphi_k\} = Aut_K E$.

57.3 Examples: (a) Consider the extension $\mathbb{C}$ over $\mathbb{R}$. Now $\mathbb{C}$ is Galois over $\mathbb{R}$ and $Aut_{\mathbb{R}}\mathbb{C} = \{\iota, \varphi\}$, where φ is the conjugation mapping. Thus $N_{\mathbb{C}/\mathbb{R}}(a + bi) = (a + bi)(a + bi)\varphi = (a + bi)(a - bi) = a^2 + b^2$ and $T_{\mathbb{C}/\mathbb{R}}(a + bi) = (a + bi) + ((a + bi)\varphi) = (a + bi) + (a - bi) = 2a$ for any $a + bi \in \mathbb{C}$ $(a, b \in \mathbb{R})$.

(b) $\mathbb{Q}(\sqrt{2})$ is a Galois extension of $\mathbb{Q}$ and $Aut_{\mathbb{Q}}\mathbb{Q}(\sqrt{2}) = \{\iota, \varphi\}$, where φ is the homomorphism $\sqrt{2} \to -\sqrt{2}$. Thus $N_{\mathbb{Q}(\sqrt{2})/\mathbb{Q}}(a + b\sqrt{2}) = (a + b\sqrt{2})(a + b\sqrt{2})\varphi = (a + b\sqrt{2})(a - b\sqrt{2}) = a^2 - 2b^2$ and $T_{\mathbb{Q}(\sqrt{2})/\mathbb{Q}}(a + b\sqrt{2}) = (a + b\sqrt{2}) + (a + b\sqrt{2})\varphi = (a + b\sqrt{2}) + (a - b\sqrt{2}) = 2a$ for any $a + b\sqrt{2} \in \mathbb{Q}(\sqrt{2})$ $(a, b \in \mathbb{Q})$.

(c) $\mathbb{Q}(\sqrt[3]{2})$ is a separable extension of $\mathbb{Q}$, but not Galois over $\mathbb{Q}$. A normal closure of $\mathbb{Q}$ over $\mathbb{Q}(\sqrt[3]{2})$ is $\mathbb{Q}(\sqrt[3]{2}, \omega)$. There are exactly three $\mathbb{Q}$-homomorphisms from $\mathbb{Q}(\sqrt[3]{2})$ into $\mathbb{Q}(\sqrt[3]{2}, \omega)$, namely $\sqrt[3]{2} \to \sqrt[3]{2}$ (the identity), $\sqrt[3]{2} \to \sqrt[3]{2}\omega$ and $\sqrt[3]{2} \to \sqrt[3]{2}\omega^2$. So $N_{\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}}(a + b\sqrt[3]{2} + c\sqrt[3]{2^2})$

$$= (a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2)(a + b\sqrt[3]{2}\omega + c(\sqrt[3]{2})^2\omega^2)(a + b\sqrt[3]{2}\omega^2 + c(\sqrt[3]{2})^2\omega)$$
$$= \cdots = a^3 + 2b^3 + 4c^3 - 2abc$$

and $T_{\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}}(a + b\sqrt[3]{2} + c\sqrt[3]{2^2})$

$$= (a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2) + (a + b\sqrt[3]{2}\omega + c(\sqrt[3]{2})^2\omega^2) + (a + b\sqrt[3]{2}\omega^2 + c(\sqrt[3]{2})^2\omega)$$

$= 3a$

for any $a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 \in \mathbb{Q}(\sqrt[3]{2})$ $(a,b,c \in \mathbb{Q})$.

In these examples, norm and trace are found to be in the base field. This is always true. In fact, the norm and trace of an element are essentially coefficients of the minimal polynomial of that element. In particular, they are independent of the normal closure that we use in their definition. We now prove these assertions.

57.4 Lemma: *Let K be a field and E a finite dimensional separable extension of K. Let a,b be arbitrary elements of E.*

(1) $N_{E/K}(ab) = N_{E/K}(a)N_{E/K}(b)$ *and* $T_{E/K}(a + b) = T_{E/K}(a) + T_{E/K}(b)$.

(2) *If* $b \in K$, *then* $N_{E/K}(b) = b^{|E:K|}$ *and* $T_{E/K}(b) = |E:K|b$.

(3) *If* $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in K[x]$ *is the minimal polynomial of* b *over* K, *then*

$$N_{E/K}(b) = ((-1)^n a_0)^{|E:K(b)|} \text{ and } T_{E/K}(b) = |E:K(b)|(-a_{n-1}).$$

Proof: Let N be a normal closure of K over E and let $\{\varphi_1,\varphi_2,\ldots,\varphi_k\}$ be the set of all K-homomorphisms from E into N. In view of the comments above, their number k is the number of roots of the minimal polynomial of a primitive element of the extension E/K, hence $k = |E:K|$ (or use Lemma 57.1 with $L = K$).

(1) Clearly
$$\begin{aligned} N_{E/K}(ab) &= (ab\varphi_1)(ab\varphi_2)\ldots(ab\varphi_k) \\ &= (a\varphi_1 b\varphi_1)(a\varphi_2 b\varphi_2)\ldots(a\varphi_k b\varphi_k) \\ &= (a\varphi_1)(b\varphi_1)(a\varphi_2)(b\varphi_2)\ldots(a\varphi_k)(b\varphi_k) \\ &= (a\varphi_1)(a\varphi_2)\ldots(a\varphi_k)(b\varphi_1)(b\varphi_2)\ldots(b\varphi_k) \\ &= N_{E/K}(a)N_{E/K}(b) \end{aligned}$$

and
$$\begin{aligned} T_{E/K}(a + b) &= (a + b)\varphi_1 + (a + b)\varphi_2 + \cdots + (a + b)\varphi_k \\ &= (a\varphi_1 + b\varphi_1) + (a\varphi_2 + b\varphi_2) + \cdots + (a\varphi_k + b\varphi_k) \\ &= a\varphi_1 + b\varphi_1 + a\varphi_2 + b\varphi_2 + \cdots + a\varphi_k + b\varphi_k \\ &= (a\varphi_1 + a\varphi_2 + \cdots + a\varphi_k) + (b\varphi_1 + b\varphi_2 + \cdots + b\varphi_k) \end{aligned}$$

$$= T_{E/K}(a) + T_{E/K}(b).$$

(2) If $b \in K$, then $b\varphi_i = b$ for all $i = 1,2,\ldots,k$ and

$$N_{E/K}(b) = (b\varphi_1)(b\varphi_2)\ldots(b\varphi_k) = bb\ldots b = b^k$$

$$T_{E/K}(b) = b\varphi_1 + b\varphi_2 + \cdots + b\varphi_k = b + b + \cdots + b = kb.$$

(3) Let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in K[x]$ be the minimal polynomial of b over K and let $b = b_1, b_2, \ldots, b_n$ be the roots of $f(x)$ in N. Then $n = |K(b):K|$ and $f(x) = (x - b_1)(x - b_2)\ldots(x - b_n)$. Thus

$$b_1 + b_2 + \cdots + b_n = -a_{n-1} \text{ and } b_1b_2\ldots b_n = (-1)^n a_0.$$

Let us write $|E: K(b)| = s$ so that $k = sn$. There are exactly n K-homomorphisms $\alpha_1, \alpha_2, \ldots, \alpha_n$ from $K(b)$ into N (namely $\alpha_i: b \to b_i$). The restriction to $K(b)$ of any φ_j $(j = 1,2,\ldots,k)$ is one of these $\alpha_1, \alpha_2, \ldots, \alpha_n$, and each α_i $(i = 1,2,\ldots,n)$ can be extended to precisely s K-homomorphisms from E into N (Lemma 57.1). Let these extensions of α_i be $\alpha_i^{(1)}, \alpha_i^{(2)}, \ldots, \alpha_i^{(s)}$. In this way, we obtain ns K-homomorphisms $\alpha_i^{(m)}: E \to N$ $(i = 1,2,\ldots,n;\ m = 1,2,\ldots,s)$. Since $ns = k$, we get

$$\{\varphi_1, \varphi_2, \ldots, \varphi_k\} = \{\alpha_i^{(m)}: i = 1,2,\ldots,n \text{ and } m = 1,2,\ldots,s\}.$$

Thus

$$N_{E/K}(b) = (b\varphi_1)(b\varphi_2)\ldots(b\varphi_k) = \prod_{i=1}^{n}\prod_{m=1}^{s} b\alpha_i^{(m)} = \prod_{i=1}^{n}(b\alpha_i)^s$$

$$= \prod_{i=1}^{n}(b_i)^s = (\prod_{i=1}^{n} b_i)^s = ((-1)^n a_0)^s$$

and $T_{E/K}(b) = b\varphi_1 + b\varphi_2 + \cdots + b\varphi_k = \sum_{i=1}^{n}\sum_{m=1}^{s} b\alpha_i^{(m)} = \sum_{i=1}^{n} s(b\alpha_i)$

$$= s\sum_{i=1}^{n} b_i = s(-a_{n-1}). \qquad \square$$

We have already mentioned that $N_{E/K}(a)$ and $T_{E/K}(a)$ depend on the fields E and K. It is clear from the definition or from Lemma 57.4(3) that $N_{E/K}(a)$ and $T_{E/K}(a)$ will be distinct from $N_{L/K}(a)$ and $T_{L/K}(a)$ and also from $N_{E/L}(a)$ and $T_{E/L}(a)$ if L is an intermediate field (with $a \in L$ in the first case).

Norm and trace behave very reasonably through intermediate fields: we have $N_{E/K} = N_{L/K} \circ N_{E/L}$ and $T_{E/K} = T_{L/K} \circ T_{E/L}$ for any intermediate field L

of E/K. This is the content of the next theorem. Although we know the structure of extensions of homomorphisms in the separable case, we give a new argument that works in more general situations.

57.5 Lemma: *Let K be a field and E a finite dimensional separable extension of K. Let a be an arbitrary element of E. Then*

$$N_{E/K}(a) = N_{L/K}(N_{E/L}(a)) \text{ and } T_{E/K}(a) = T_{L/K}(T_{E/L}(a)).$$

Proof: (The assertion is meaningful, for $N_{E/L}(a)$ is an element of L by Lemma 57.4(3), thus we can take the norm of $N_{L/K}(a) \in L$ over K. The claim is that this is equal to the norm of $a \in E$ over K. Similarly for the trace.)

$$\begin{array}{ccc} & E & \\ s & \downarrow N_{E/L} & \\ N_{E/K} & L & \\ n & \downarrow N_{L/K} & \\ & K & \end{array}$$

The proof has been foreshadowed is Lemma 57.4. Let N be a normal closure of K over E. Then N is Galois over K by Theorem 55.11 and N is Galois over L by Theorem 54.25(1). We choose a field M such that (i) $E \subseteq M \subseteq N$; (ii) M is Galois over L; (iii) A is not Galois over L for any field A with $E \subseteq A \subset M$. This is possible because N/K is Galois, N/E is also Galois and so N/E is separable (Theorem 55.10) and there are only finitely many intermediate fields of N/E (Theorem 55.15). M is a normal closure of L over E. Likewise, we choose a field R such that (i) $L \subseteq R \subseteq N$; (ii) R is Galois over K; (iii) A is not Galois over L for any field A with $L \subseteq A \subset R$. Thus R is a normal closure of K over L.

We put $|E:K| = k$, $|E:L| = s$ and $|L:K| = n$ so that $k = sn$. Let

$\{\varphi_1, \varphi_2, \ldots, \varphi_k\}$ be the set of all K-homomorphisms from E into N,
$\{\beta_1, \beta_2, \ldots, \beta_s\}$ the set of all L-homomorphisms from E into $M \subseteq N$ and
$\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ the set of all K-homomorphisms from L into $R \subseteq N$.

Then
$$N_{E/K}(a) = (a\varphi_1)(a\varphi_2)\ldots(a\varphi_k),$$
$$N_{E/L}(a) = (a\beta_1)(a\beta_2)\ldots(a\beta_s),$$

$$N_{L/K}(b) = (b\alpha_1)(b\alpha_2)\dots(b\alpha_n)$$

for any $a \in E$, $b \in L$.

N is a splitting field of a polynomial $f(x) \in K[x]$ over K (Theorem 55.11 and Theorem 55.7) and therefore N is a splitting field of $f(x)$ over L and over $L\alpha_i$ (Example 53.5(e)). The isomorphism $\alpha_i: L \to L\alpha_i$ ($\subseteq N$) can be extended to an isomorphism $\alpha_i^{(1)}: N \to N$ (Theorem 53.7). Here of course $\alpha_i^{(1)}: N \to N$ is a K-homomorphism.

We claim $\{\varphi_1,\varphi_2,\dots,\varphi_k\} = \{\beta_j\,\alpha_i^{(1)}: i = 1,2,\dots,n; j = 1,2,\dots,s\}$. Since $k = ns$, we must merely show that $\beta_{j'}\alpha_{i'}^{(1)} \neq \beta_j\,\alpha_i^{(1)}$ when $(i,j) \neq (i',j')$. Indeed, if $\beta_{j'}\alpha_{i'}^{(1)} = \beta_j\,\alpha_i^{(1)}$, then the restriction of $\beta_{j'}\alpha_{i'}^{(1)}$ and $\beta_j\,\alpha_i^{(1)}$ to L must be equal and since β_j and $\beta_{j'}$ fix each element in L, we get $\alpha_{i'}^{(1)}{}_{|L} = \alpha_i^{(1)}{}_{|L}$, so $\alpha_{i'} = \alpha_i$ and $i' = i$. Then, as $\alpha_i^{(1)}$ is one-to-one, $i' = i$ and $\beta_{j'}\alpha_{i'}^{(1)} = \beta_j\,\alpha_i^{(1)}$ imply that $\beta_{j'} = \beta_j$ and $j' = j$. This establishes the claim.

Thus, since $N_{E/L}(a) \in L$ by Lemma 57.5(3), we get

$$\begin{aligned} N_{E/K}(a) &= (a\varphi_1)(a\varphi_2)\dots(a\varphi_k) = \prod_{i=1}^{n}\prod_{j=1}^{s} a\beta_j\,\alpha_i^{(1)} \\ &= \prod_{i=1}^{n}\Big(\prod_{j=1}^{s} a\beta_j\Big)\alpha_i^{(1)} = \prod_{i=1}^{n}(N_{E/L}(a))\alpha_i^{(1)} = \prod_{i=1}^{n}(N_{E/L}(a))\alpha_i \\ &= N_{L/K}(N_{E/L}(a)) \end{aligned}$$

and similarly $T_{E/K}(a) = T_{L/K}(T_{E/L}(a))$. □

57.6 Definition: Let E be a field and let $\{\varphi_1,\varphi_2,\dots,\varphi_k\}$ be a finite set of field automorphisms of E. If, for any $a_1,a_2,\dots,a_k \in E$,

$$a_1(b\varphi_1) + a_2(b\varphi_2) + \dots + a_k(b\varphi_k) = 0 \text{ for all } b \in E$$

implies $a_1 = a_2 = \dots = a_k = 0$, then $\{\varphi_1,\varphi_2,\dots,\varphi_k\}$ is said to be *linearly independent*.

Equivalently, $\{\varphi_1,\varphi_2,\dots,\varphi_k\}$ is linearly independent provided, for each k-tuple $(a_1,a_2,\dots,a_k)$ of elements from E, where at least one a_i is distinct from 0, there is a $b \in E$ such that

$$a_1(b\varphi_1) + a_2(b\varphi_2) + \dots + a_k(b\varphi_k) \neq 0.$$

57.7 Lemma: *Let E be a field and let $\{\varphi_1,\varphi_2,\ldots,\varphi_k\}$ be a finite set of field automorphisms of E. If $\varphi_1,\varphi_2,\ldots,\varphi_k$ are pairwise distinct, then $\{\varphi_1,\varphi_2,\ldots,\varphi_k\}$ is linearly independent.*

Proof: (cf. Lemma 54.15; note that we do not assume $\{\varphi_1,\varphi_2,\ldots,\varphi_k\}$ is a group.) Suppose, by way of contradiction, $\varphi_1,\varphi_2,\ldots,\varphi_k$ are distinct automorphisms of E and that $\{\varphi_1,\varphi_2,\ldots,\varphi_k\}$ is not linearly independent. Then there are elements $a_1,a_2,\ldots,a_k$, not all zero, in E such that

$$a_1(b\varphi_1) + a_2(b\varphi_2) + \cdots + a_k(b\varphi_k) = 0 \qquad \text{for all } b \in E. \tag{1}$$

Let r be the smallest number of nonzero components a_i in all the k-tuples $(a_1,a_2,\ldots,a_k) \in E\times E\times\ldots\times E\setminus\{(0,0,\ldots,0)\}$ satisfying (1) and choose a k-tuple $(c_1,c_2,\ldots,c_k)$ with exactly r nonzero components. We have $r > 1$. Renumbering the automorphisms, we may assume $c_1,\ldots,c_r$ are distinct from zero and (in case $r < k$) $c_{r+1} = \cdots = c_k = 0$.

Then $$c_1(b\varphi_1) + c_2(b\varphi_2) + \cdots + c_r(b\varphi_r) = 0 \quad \text{for all } b \in E. \tag{2}$$

Since $\varphi_1,\varphi_2,\ldots,\varphi_k$ are distinct, $\varphi_1 \neq \varphi_2$ and there is a $u \in E$ with $u\varphi_1 \neq u\varphi_2$. Writing ub in place of b in (2) and using $(ub)\varphi_i = u\varphi_i\cdot u\varphi_i$, we get

$$c_1(u\varphi_1)(b\varphi_1) + c_2(u\varphi_2)(b\varphi_2) + \cdots + c_r(u\varphi_r)(b\varphi_r) = 0 \qquad \text{for all } b \in E. \tag{3}$$

Multiplying (2) by $u\varphi_1$, we obtain

$$c_1(u\varphi_1)(b\varphi_1) + c_2(u\varphi_1)(b\varphi_2) + \cdots + c_r(u\varphi_1)(b\varphi_r) = 0 \qquad \text{for all } b \in E. \tag{4}$$

Subtraction gives

$$[c_2(u\varphi_2 - u\varphi_1)](b\varphi_2) + \cdots + [c_r(u\varphi_r - u\varphi_1)](b\varphi_r) = 0 \qquad \text{for all } b \in E.$$

where at least $c_2(u\varphi_2 - u\varphi_1) \neq 0$. Hence there is a k-tuple

$$(0,c_2(u\varphi_2 - u\varphi_1),\ldots,c_r(u\varphi_r - u\varphi_1),0,\ldots,0) \neq (0,0,\ldots,0)$$

with at most $r-1$ nonzero components satisfying (1), contrary to the definition of r. Therefore $\{\varphi_1,\varphi_2,\ldots,\varphi_k\}$ is linearly independent. □

We now characterize all elements with trace 0 and all elements with norm 1 in case of a Galois extension with a finite cyclic group. The second part of Theorem 57.9 (formulated for finite dimensional extensions of $\mathbb{Q}$) is the theorem with number 90 in D. Hilbert's (1862-1943) famous report on algebraic number theory and is known as "Hilbert's theorem 90". It is the beginning of cohomology theory.

57.8 Definition: Let E/K be a field extension. If E is algebraic and Galois over K, and if the Galois group $Aut_K E$ is cyclic, then E is called a *cyclic extension of* K and E/K is said to be *cyclic*.

57.9 Theorem: *Let E/K be a finite dimensional cyclic extension and let σ be a generator of $Aut_K E$. Let $a \in E$.*
(1) $T_{E/K}(a) = 0$ *if and only if there is an element* $b \in E$ *with* $a = b - b\sigma$.
(2) $N_{E/K}(a) = 1$ *if and only if there is an element* $b \in E\backslash\{0\}$ *with* $a = b/b\sigma$.

Proof: Let $|E:K| = n$. Then $|Aut_K E| = n$ by the fundamental theorem of Galois theory and $Aut_K E = \{1,\sigma,\sigma^2,\ldots,\sigma^{n-2},\sigma^{n-1}\}$, with $\sigma^n = 1$ and $o(\sigma) = n$. For convenience, we write T instead of $T_{E/K}$ and N instead of $N_{E/K}$.

(1) If $a = b - b\sigma$, then we have a telescoping sum:
$T(a) = a + a\sigma + a\sigma^2 + \cdots + a\sigma^{n-2} + a\sigma^{n-1}$
$= (b - b\sigma) + (b - b\sigma)\sigma + (b - b\sigma)\sigma^2 + \cdots + (b - b\sigma)\sigma^{n-2} + (b - b\sigma)\sigma^{n-1}$
$= (b - b\sigma) + (b\sigma - b\sigma^2) + (b\sigma^2 - b\sigma^3) + \cdots + (b\sigma^{n-2} - b\sigma^{n-1}) + (b\sigma^{n-1} - b\sigma^n)$
$= b - b\sigma^n = b - b = 0.$

Conversely, assume that $T(a) = 0$. We first find an element c in E with $T(c) = 1$. Since $o(\sigma) = n$, the automorphisms $1,\sigma,\sigma^2,\ldots,\sigma^{n-2},\sigma^{n-1}$ are distinct and $\{1,\sigma,\sigma^2,\ldots,\sigma^{n-2},\sigma^{n-1}\}$ is linearly independent by Lemma 57.7. So there is a $u \in E$ with

$$T(u) = u + u\sigma + u\sigma^2 + \cdots + u\sigma^{n-2} + u\sigma^{n-1} \neq 0.$$

Let $c = u/T(u)$. Since $T(u) \in K$ and E is Galois over K, we have $(T(u))\sigma^j = T(u)$ for any $j = 0,1,2,\ldots,n-1$ and thus

$$T(c) = \frac{u}{T(u)} + (\frac{u}{T(u)})\sigma + (\frac{u}{T(u)})\sigma^2 + \cdots + (\frac{u}{T(u)})\sigma^{n-2} + (\frac{u}{T(u)})\sigma^{n-1}$$

$$= \frac{u}{T(u)} + \left(\frac{u\sigma}{T(u)\sigma}\right) + \left(\frac{u\sigma^2}{T(u)\sigma^2}\right) + \cdots + \left(\frac{u\sigma^{n-2}}{T(u)\sigma^{n-2}}\right) + \left(\frac{u\sigma^{n-1}}{T(u)\sigma^{n-1}}\right)$$
$$= \frac{u}{T(u)} + \frac{u\sigma}{T(u)} + \frac{u\sigma^2}{T(u)} + \cdots + \frac{u\sigma^{n-2}}{T(u)} + \frac{u\sigma^{n-1}}{T(u)}$$
$$= \frac{T(u)}{T(u)} = 1.$$

We put $b = ac + (a + a\sigma)(c\sigma) + (a + a\sigma + a\sigma^2)(c\sigma^2) + (a + a\sigma + a\sigma^2 + a\sigma^3)(c\sigma^3)$
$+ \cdots + (a + a\sigma + a\sigma^2 + \cdots + a\sigma^{n-2})(c\sigma^{n-2}) \in E.$

Then

$$\begin{aligned} b - b\sigma = {} & ac + (a + a\sigma)(c\sigma) + (a + a\sigma + a\sigma^2)(c\sigma^2) \\ & + (a + a\sigma + a\sigma^2 + a\sigma^3)(c\sigma^3) \\ & + \cdots + (a + a\sigma + a\sigma^2 + \cdots + a\sigma^{n-2})(c\sigma^{n-2}). \\ & - (a\sigma)(c\sigma) - (a\sigma + a\sigma^2)(c\sigma^2) - (a\sigma + a\sigma^2 + a\sigma^3)(c\sigma^3) \\ & - (a\sigma + a\sigma^2 + a\sigma^3 + a\sigma^4)(c\sigma^4) \\ & - \cdots - (a\sigma + a\sigma^2 + a\sigma^3 + \cdots + a\sigma^{n-1})(c\sigma^{n-1}) \\ = {} & ac + a(c\sigma) + a(c\sigma^2) + \cdots + a(c\sigma^{n-2}) - (a\sigma + a\sigma^2 + \cdots + a\sigma^{n-1})(c\sigma^{n-1}) \\ = {} & ac + a(c\sigma) + a(c\sigma^2) + \cdots + a(c\sigma^{n-2}) - (T(a) - a)(c\sigma^{n-1}) \\ = {} & ac + a(c\sigma) + a(c\sigma^2) + a(c\sigma^3) + \cdots + a(c\sigma^{n-1}) = aT(c) = a. \end{aligned}$$

Hence $a = b - b\sigma$ for some $b \in E$ when $T(a) = 0$.

(2) If $a = b/b\sigma$ for some $b \in E\backslash\{0\}$, then

$$N(a) = \frac{b}{b\sigma}\cdot\left(\frac{b}{b\sigma}\right)\sigma\cdot\left(\frac{b}{b\sigma}\right)\sigma^2 \cdots \left(\frac{b}{b\sigma}\right)\sigma^{n-2}\cdot\left(\frac{b}{b\sigma}\right)\sigma^{n-1}$$
$$= \frac{b}{b\sigma}\cdot\frac{b\sigma}{b\sigma^2}\cdot\frac{b\sigma^2}{b\sigma^3}\cdots\frac{b\sigma^{n-2}}{b\sigma^{n-1}}\cdot\frac{b\sigma^{n-1}}{b\sigma^n} = \frac{b}{b} = 1.$$

Conversely, assume $N(a) = 1$. Then of course $a \neq 0$. From Lemma 57.7, it follows that there is a $d \in E$ for which

$$b := (a)d + (a\cdot a\sigma)d\sigma + (a\cdot a\sigma\cdot a\sigma^2)d\sigma^2 + \cdots + (a\cdot a\sigma\cdot a\sigma^2 \ldots a\sigma^{n-2})d\sigma^{n-2}$$
$$+ (a\cdot a\sigma\cdot a\sigma^2 \ldots a\sigma^{n-2}\cdot a\sigma^{n-1})d\sigma^{n-1}$$

is distinct from 0. Then we get

$$\begin{aligned} a(b\sigma) = {} & a(a\sigma)d\sigma + a(a\sigma\cdot a\sigma^2)d\sigma^2 + a(a\sigma\cdot a\sigma^2\cdot a\sigma^3)d\sigma^3 \\ & + \cdots + a(a\sigma\cdot a\sigma^2\cdot a\sigma^3 \ldots a\sigma^{n-1})d\sigma^{n-1} + a(a\sigma\cdot a\sigma^2\cdot a\sigma^3 \ldots a\sigma^{n-1}\cdot a\sigma^n)d\sigma^n \\ = {} & b - (a)d + a\cdot N(a)\cdot d\sigma^n = b - ad + a\cdot 1\cdot d = b. \end{aligned}$$

Since $b \neq 0$, also $b\sigma \neq 0$ and $a(b\sigma) = b$ gives $a = b/b\sigma$. □

We close this paragraph with two applications of Theorem 57.9. We describe cyclic extensions. The degree is the characteristic in the first case and relatively prime to the characteristic in the second case.

57.10 Theorem: *Let K be a field of characteristic $p \neq 0$ and let E be a cyclic extension of K with $|E:K| = p$. Then there is an $a \in K$ such that $f(x) = x^p - x - a \in K[x]$ is irreducible in $K[x]$ and $E = K(t)$ for any root t of $f(x)$.*

Proof: By hypothesis, E is Galois over K and $Aut_K E$ is cyclic of order p, say $Aut_K E = \langle \sigma \rangle$. Then $T_{E/K}(1) = 1 + 1\sigma + 1\sigma^2 + \ldots + 1\sigma^{p-1} = p = 0$, so $1 = b - b\sigma$ for some $b \in E$ (Theorem 57.9(1)). Let $u = -b$. Then $u\sigma = 1 + u$ and we get $u^p\sigma = (u\sigma)^p = (1 + u)^p = 1^p + u^p = 1 + u^p$. Hence

$$(u^p - u)\sigma = u^p\sigma - u\sigma = (1 + u^p) - (1 + u) = u^p - u$$

and $u^p - u$ is fixed by σ and thus by all automorphisms in $Aut_K E$. Since E is Galois over K, this gives $u^p - u \in K$. Let us put $u^p - u = a$. Thus u is a root of $f(x) = x^p - x - a \in K[x]$.

It remains to show that $f(x)$ is irreducible over K and that $E = K(t)$ for any root t of $f(x)$. Since b is not fixed by σ, we see $b \notin K$, so $u \notin K$ and thus $K \subset K(u) \subseteq E$. But $|E:K| = p$ is prime and so there is no intermediate field of E/K distinct from K and E. This forces $K(u) = E$. Then $\deg f(x) = p = |E:K| = |K(u):K|$ = degree of the minimal polynomial of u over K. Since the minimal polynomial of u over K divides $f(x)$, we deduce that $f(x)$ is the minimal polynomial of u over K. In particular, $f(x)$ is irreducible in $K[x]$.

Now for any $j \in \mathbb{F}_p \subseteq K$, there holds $j^p = j$ and consequently

$$f(u + j) = (u + j)^p - (u + j) - a = u^p + j^p - u - j - a = u^p - u - a = 0.$$

So $u, u + 1, u + 2, \ldots, u + p - 1 \in E$ are roots of $f(x)$. Since $f(x)$ has p roots, any root t of $f(x)$ is equal to $u + j$ for some $j \in \mathbb{F}_p$. So we get $K(t) = K(u + j) = K(u) = E$ for any root t of $f(x)$. □

57.11 Theorem: *Let K be a field and let E be a cyclic extension of K of degree $|E:K| = n$. Assume that either char $K = 0$ or char $K \neq 0$ but char K does not divide n. Assume, in addition, that $x^n - 1$ splits in K. Then there is an $a \in K$ such that $f(x) = x^n - a \in K[x]$ is irreducible in $K[x]$ and $E = K(u)$ for any root u of $f(x)$.*

Proof: By hypothesis, $Aut_K E$ is a cyclic group, say $Aut_K E = <\sigma>$, and $o(\sigma) = |Aut_K E| = |E:K| = n$. All roots of the polynomial $x^n - 1$, which splits in K, are simple since its derivative $nx^{n-1} \neq 0$ in view of the assumption on *char* K. Thus there are exactly n distinct roots of $x^n - 1$ in K. Since $r^n = s^n = 1$ implies $(rs)^n = 1$, the roots of $x^n - 1$ make up a subgroup of $K^\times$. Any finite subgroup of $K^\times$ is cyclic (Theorem 52.18), so the roots of $x^n - 1$ form a cyclic group of order n. Let $r \in K$ be a generator of this group so that the n roots of $x^n - 1$ are $1, r, r^2, \ldots, r^{n-1}$.

We have $N_{E/K}(r) = r^n = 1$ (Lemma 57.4(2)) and so there is a $b \in E$ with $r = b/b\sigma$ (Theorem 57.9(2)). Let $u = 1/b$. Then $u \in E\backslash\{0\}$ and $u\sigma = ur$. This implies $u^n\sigma = (u\sigma)^n = (ur)^n = u^n r^n = u$, so u^n is fixed by σ, so by $Aut_K E$, and therefore $u^n \in K$. Let us put $u^n = a$.

Then $x^n - a \in K[x]$ and this polynomial has n roots $u, ur, ur^2, \ldots, ur^{n-1}$ in $K(u)$, which are all distinct. So $x^n - a$ splits in $K(u)$, but not in a proper subfield of $K(u)$ containing K, since any intermediate field of $K(u)/K$, in which $x^n - a$ splits, must contain the root u and hence must be identical with $K(u)$. Thus $K(u)$ is a splitting field of $x^n - a$ over K. Since the roots of $x^n - a$ are distinct, the irreducible factors of $x^n - a$ are separable over K and thus $K(u)$ is Galois over K (Theorem 55.7). In particular, $|Aut_K K(u)| = |K(u):K|$.

Any K-automorphism $\sigma^j \in Aut_K E$ $(j = 0,1,2,\ldots,n-1)$ sends u to $ur^j \in K(u)$, thus the restriction of σ^j to $K(u)$ is a K-automorphism of $K(u)$ (Theorem 42.22). Since $u\sigma^i = ur^i \neq ur^j = u\sigma^j$ when $i,j \in \{0,1,2,\ldots,n-1\}$ and $i \neq j$, we see that these K-automorphisms of $K(u)$ are distinct. Hence there are at least n K-automorphisms of $K(u)$. This implies $|K(u):K| = |Aut_K K(u)| \geqslant n$. From $n = |E:K| \geqslant |K(u):K| \geqslant n$, we get $|K(u):K| = n$, whence $E = K(u)$.

Finally, since the minimal polynomial of u over K divides $x^n - a$ and $deg(x^n - a) = n = |K(u):K|$ = degree of the minimal polynomial of u over K, we deduce that $x^n - a$ is the minimal polynomial of u over K and $x^n - a$ is irreducible in $K[x]$. Moreover, any root t of $x^n - a$ is equal to ur^j for some $j = 0,1,2,\ldots,n-1$ and, since $r \in K$, we get $K(t) = K(ur^j) = K(u) = E$ for any root t of $x^n - a$. □

Exercises

1. Let K be a field and let E be a finite dimensional separable extension of K. Prove that, for any $k \in K$, there is an $a \in E$ such that $T_{E/K}(a) = k$.

2. Let $K \subseteq L \subseteq E \subseteq N$ be fields and assume that N is normal over K. If s is the cardinal number of L-homomorphisms from E into N and n is the cardinal number of K-homomorphisms from L into N, prove that sn is the cardinal number of K-homomorphisms from E into N.

3. Let K be a field of characteristic $p \neq 0$ and $f(x) = x^p - x - a \in K[x]$. Show that $f(x)$ either splits in K or is irreducible in $K[x]$.

4. Let K be a field of characteristic $p \neq 0$ and $f(x) = x^p - x - a \in K[x]$. Prove that if $f(x)$ is irreducible in $K[x]$ and u a root of $f(x)$, then $K(u)$ is a cyclic extension of K of degree p.

5. Let K be a field and $n \in \mathbb{N}$. Assume that either *char* $K = 0$ or *char* $K \neq 0$ but *char* K does not divide n. Assume that $x^n - 1$ splits in K. Prove that, if $a \in K$ and u a root of $f(x) = x^n - a \in K[x]$, then $K(u)$ is a cyclic extension of K and $|K(u):K|$ divides n and $u^{|K(u):K|} \in K$.

§58
Cyclotomic Fields

The theory of cyclotomy is concerned with the problem of dividing the perimeter of a circle into a given number of equal parts (cyclo-tomy means: circle-division). Consider the unit circle in the complex plane. The points dividing this unit circle into n equal parts are the points $e^{2\pi i/n} = \cos(2\pi/n) + i\sin(2\pi/n)$ and the geometric problem of cyclotomy is equivalent to studying the fields $\mathbb{Q}(e^{2\pi i/n}) \subseteq \mathbb{C}$. The complex numbers $e^{2\pi i/n}$ are roots of the polynomial $x^n - 1$ and $\mathbb{Q}(e^{2\pi i/n})$ is a splitting field of $x^n - 1$. The splitting fields of such polynomials over any field K will be called cyclotomic fields (although they may not be relevant to the geometric problem of circle division).

58.1 Definition: Let K be a field and $1 \in K$ the identity element of K. Let $n \in \mathbb{N}$. An extension field E of K is called a *cyclotomic extension of K (of order n)* if E is a splitting field of $x^n - 1 \in K[x]$ over K.

58.2 Definition: Let K be a field. A root of the polynomial $x^n - 1 \in K[x]$ is called an *n-th root of unity* or, if there is no need to be exact, simply a *root of unity*.

58.3 Lemma: *Let K be a field of characteristic $p \neq 0$ and let $n \in \mathbb{N}$, where $n = p^a m$ and $(p,m) = 1$. Let u be an element in an extension field of K. Then u is an n-th root of unity if and only if u is an m-th root of unity.*

Proof: If u is an m-th root of unity, then $u^m = 1$, so $u^n = (u^m)^{p^a} = 1^{p^a} = 1$ and u is an n-th root of unity. If u is an n-th root of unity, then $0 = u^n - 1 = (u^m)^{p^a} - 1 = (u^m - 1)^{p^a}$, so $u^m - 1 = 0$ and u is an m-th root of unity. □

So in the situation of Lemma 58.3, a splitting field of $x^n - 1$ over K is also a splitting field of $x^m - 1$ over K, and conversely. For this reason, in case *char* $K \neq 0$, it is no loss of generality to assume that the order of a cyclotomic extension is relatively prime to the characteristic of K.

58.4 Lemma: *Let K be a field and E an extension field of K containing all n-th roots of unity. Assume char $K = 0$ or (char K,n) = 1. Then the set of all n-th roots of unity is a cyclic group of order n under multiplication.*

Proof: If u and t are n-th roots of unity, then $(ut)^n = u^n t^n = 1 \cdot 1 = 1$ and ut is also an n-th root of unity. Since the number of n-th roots of unity is at most n (Theorem 35.7), it follows that the set of all n-th roots of unity is a subgroup of $K^\times$ (Lemma 9.3(1)). This group of n-th roots of unity is cyclic by Theorem 52.18. To prove that the order of this group is equal to n, we must only show that all roots of $x^n - 1$ are simple. This follows from the fact that the derivative nx^{n-1} of $x^n - 1$ is distinct from zero (because of the assumption *char* $K = 0$ or (*char* K,n) = 1) so that $x^n - 1$ and nx^{n-1} have no common root. □

58.5 Definition: Let K be a field and E an extension field of K containing all n-th roots of unity. Assume *char* $K = 0$ or (*char* K,n) = 1. A generator of the cyclic group of all n-th roots of unity is called a *primitive n-th root of unity.*

ζ is a primitive n-th root of unity if and only if $o(\zeta) = n$. If ζ is a primitive n-th root of unity, then all n-th roots of unity are given without duplication in the list

$$1 = \zeta^0, \zeta^1, \zeta^2, \zeta^3, \ldots, \zeta^{n-1}$$

or in the list

$$\zeta, \zeta^2, \zeta^3, \ldots, \zeta^{n-1}, \zeta^n = 1$$

and ζ^j has order $n/(n,j)$ (Lemma 11.9(2)). Hence ζ^j is a primitive n-th root of unity if and only if $(n,j) = 1$. There are therefore $\varphi(n)$ primitive n-th roots of unity (cf. §11).

If u is a root of unity and $o(u) = d$, then, by definition, u is a primitive d-th root of unity.

1 is a primitive first root of unity, $-1 \in \mathbb{C}$ is a primitive second root of unity, $\omega \in \mathbb{C}$ and ω^2 are primitive third roots of unity, $i \in \mathbb{C}$ and $-i \in \mathbb{C}$ are primitive fourth roots of unity.

58.6 Definition: Let K be a field and $n \in \mathbb{N}$. Assume that *char* $K = 0$ or (*char* K,n) = 1. Let ζ be a primitive n-th root of unity and

$$\{\zeta_1,\zeta_2,\ldots\zeta_{\varphi(n)}\} = \{\zeta^j : j = 1,2,\ldots,n \text{ and } (n,j) = 1\}$$

the set of all primitive n-th roots of unity in some extension field of K. The monic polynomial

$$(x - \zeta_1)(x - \zeta_2)\ldots(x - \zeta_{\varphi(n)})$$

of degree $\varphi(n)$ is called the *n-th cyclotomic polynomial over* K and is denoted by $\Phi_n(x)$.

For example, over $\mathbb{Q}$, the first few cyclotomic polynomials are

$$\Phi_1(x) = x - 1, \qquad \Phi_2(x) = x - (-1) = x + 1,$$
$$\Phi_3(x) = (x - \omega)(x - \omega^2) = x^2 + x + 1, \qquad \Phi_4(x) = (x - i)(x + i) = x^2 + 1.$$

We see that these are in fact polynomials in $\mathbb{Z}[x]$. This is true for any cyclotomic polynomial. The n-th cyclotomic polynomial over K does not depend on the extension field of K in which the primitive n-th roots of unity are assumed to lie. In fact, it does not even depend on K (but only on *char* K).

58.7 Lemma: *Let K be a field, $n \in \mathbb{N}$ and assume that char $K = 0$ or (char K,n) = 1. Then*

(1) $x^n - 1 = \prod_{d|n} \Phi_d(x)$.

(2) $\Phi_n(x) \in \mathbb{Z}[x]$ *if char* $K = 0$ *and* $\Phi_n(x) \in \mathbb{Z}_p[x] = \mathbb{F}_p[x]$ *if char* $K = p \neq 0$.

Proof: (1) Any root u of $x^n - 1$ is an n-th root of unity and $o(u) = d$ for some divisor of n. Then u is a primitive d-th root of unity. Conversely, if $d|n$, any primitive d-th root of unity is an n-th root of unity with $o(u) =$

d. Thus $\Phi_d(x) = \prod_{\substack{u^n=1 \\ o(u)=d}} (x - u)$. Collecting together the roots of $x^n - 1$ with order d, for each divisor d of n, we get

$$x^n - 1 = \prod_{u^n=1} (x - u) = \prod_{d|n} \prod_{\substack{u^n=1 \\ o(u)=d}} (x - u) = \prod_{d|n} \Phi_d(x).$$

(2) Let $D = \mathbb{Z}$ in case *char* $K = 0$ and $D = \mathbb{Z}_p = \mathbb{F}_p$ in case *char* $K = p \neq 0$. We prove $\Phi_n(x) \in D[x]$ by induction on n. Since $\Phi_1(x) = x - 1$ and $\Phi_2(x) = x + 1$, we have $\Phi_n(x) \in D[x]$ when $n = 1,2$.

Suppose now $n \geqslant 3$ and that $\Phi_d(x) \in D[x]$ for all $d = 1,2,\ldots,n - 1$. From (1), we have $x^n - 1 = \Phi_n(x) \prod_{\substack{d|n \\ d\neq n}} \Phi_d(x)$. Let us put $f(x) = \prod_{\substack{d|n \\ d\neq n}} \Phi_d(x)$. Then $f(x)$ is a monic polynomial and $f(x) \in D[x]$ since, by induction, $\Phi_d(x) \in D[x]$ for all divisors d of n which are distinct from n. As $x^n - 1 \in D[x]$ and $f(x)$ is monic, there are unique polynomials $q(x)$ and $r(x)$ in $D[x]$ such that

$$x^n - 1 = q(x)f(x) + r(x), \qquad r(x) = 0 \text{ or } deg\ r(x) < deg\ f(x)$$

(Theorem 34.4). Now let E be an extension field of K containing all roots of $x^n - 1$. The division algorithm in $E[x]$ reads

$$x^n - 1 = \Phi_n(x)f(x) + 0.$$

Since $D \subseteq K \subseteq E$ and the quotient and remainder are uniquely determined, the unique quotient $q(x)$ in $D[x]$ must be the unique quotient $\Phi_n(x)$ in $E[x]$ and the unique remainder $r(x)$ in $D[x]$ must be the unique remainder 0 in $E[x]$. Hence $\Phi_n(x) = q(x) \in D[x]$. This completes the proof. □

The equation $\Phi_n(x) = \dfrac{x^n - 1}{\prod_{\substack{d|n \\ d\neq n}} \Phi_d(x)}$ is a recursive formula for $\Phi_n(x)$. Thus

$$\begin{aligned}\Phi_6(x) &= x^6 - 1/\Phi_1(x)\Phi_2(x)\Phi_3(x) \\ &= x^6 - 1/(x - 1)(x + 1)(x^2 + x + 1) = x^2 - x + 1.\end{aligned}$$

Another recursive formula is for $\Phi_n(x)$ is given in the next lemma.

58.8 Lemma: *Let K be a field, $n \in \mathbb{N}$ and assume that char $K = 0$ or (char K,n) = 1. Then*

$$\Phi_n(x) = \prod_{d|n} (x^d - 1)^{\mu(n/d)} = \prod_{d|n} (x^{n/d} - 1)^{\mu(d)}.$$

Proof: This follows immediately from Lemma 58.7(1) and Lemma 52.14 (in Lemma 52.14, let the field be $K(x)$ and let the function $F: \mathbb{N} \to K(x)^\times$ be $n \to \Phi_n(x)$). □

For example, we have, over $\mathbb{Q}$:

$$\begin{aligned}\Phi_{12}(x) &= (x^{12} - 1)^{\mu(1)}(x^6 - 1)^{\mu(2)}(x^4 - 1)^{\mu(3)}(x^3 - 1)^{\mu(4)}(x^2 - 1)^{\mu(6)}(x - 1)^{\mu(12)} \\ &= (x^{12} - 1)(x^2 - 1)/(x^6 - 1)(x^4 - 1) = x^4 - x^2 + 1,\end{aligned}$$

$$\begin{aligned}\Phi_{15}(x) &= (x^{15} - 1)^{\mu(1)}(x^5 - 1)^{\mu(3)}(x^3 - 1)^{\mu(5)}(x - 1)^{\mu(15)} \\ &= (x^{15} - 1)(x - 1)/(x^5 - 1)(x^3 - 1) \\ &= x^8 - x^7 + x^5 - x^4 + x^3 - x + 1.\end{aligned}$$

58.10 Theorem: *Let K be a field, $n \in \mathbb{N}$ and assume that char $K = 0$ or (char K,n) = 1. Let E be a cyclotomic extension of order n and let $\zeta \in E$ be a primitive n-th root of unity and let $f(x) \in K[x]$ be the minimal polynomial of ζ over K. Then*

(1) $E = K(\zeta)$;

(2) *E is Galois over K;*

(3) *$|Aut_K E|$ divides $\varphi(n)$ and $Aut_K E$ is isomorphic to a subgroup of $\mathbb{Z}_n^\times$;*

(4) $Aut_K E \cong \mathbb{Z}_n^\times \iff |Aut_K E| = \varphi(n) \iff f(x) = \Phi_n(x)$
$\iff \Phi_n(x)$ is *irreducible in* $K[x]$.

Proof: (1) Let $a_1, a_2, \ldots, a_k$ be the natural numbers less than n and relatively prime to n (where $k = \varphi(n)$) so that $\zeta^{a_1}, \zeta^{a_2}, \ldots \zeta^{a_k}$ are the roots of $\Phi_n(x)$. Now E is a splitting field of $\Phi_n(x)$ by definition, so E is generated by the roots of $\Phi_n(x)$ over K (Example 53.5(d)) and $E = K(\zeta^{a_1}, \zeta^{a_2}, \ldots \zeta^{a_k}) = K(\zeta)$.

(2) The roots of $\Phi_n(x)$ are simple because $\Phi_n(x)$ is a divisor of $x^n - 1$ and the roots of $x^n - 1$ are simple (the derivative of $x^n - 1$, being distinct from 0 since *char* $K = 0$ or (*char* K,n) $= 1$, is relatively prime to $x^n - 1$). So the irreducible factors of $\Phi_n(x)$ are separable over K. Since E is a splitting field of $\Phi_n(x)$, Theorem 55.7 shows that E is Galois over K.

(3) Since ζ is a root of $\Phi_n(x) \in K[x]$ and $f(x)$ is the minimal polynomial of ζ over K, we see $f(x)$ divides $\Phi_n(x)$ in $K[x]$ and the roots of $f(x)$ are certain of the roots of $\Phi_n(x)$. Let *deg* $f(x) = s$ and $\zeta^{m_1}, \zeta^{m_2}, \ldots \zeta^{m_s}$ be the roots of $f(x)$, where $m_1, m_2, \ldots, m_s$ are some suitable natural numbers relatively prime to n and less than n and $m_1 = 1$, say. Thus

$$f(x) = (x - \zeta^{m_1})(x - \zeta^{m_2})\ldots(x - \zeta^{m_s}).$$

Here we have $|Aut_K E| = |E:K| = |K(\zeta):K| = deg\ f(x) = s$ because E is Galois over K. Any K-automorphism of E maps ζ to one of $\zeta^{m_1}, \zeta^{m_2}, \ldots \zeta^{m_s}$. Let α_{m_i} be the K-automorphism $\zeta \to \zeta^{m_i}$ $(i = 1,2, \ldots, s)$. Since

$$\alpha_{m_i} = \alpha_{m_j} \iff \zeta^{m_i} = \zeta^{m_j} \iff m_i \equiv m_j \pmod{n} \iff i = j,$$

$\alpha_{m_1}, \alpha_{m_2}, \ldots, \alpha_{m_s}$ are pairwise distinct and $Aut_K E = \{\alpha_{m_1}, \alpha_{m_2}, \ldots, \alpha_{m_s}\}$.

Let m_i^* be the residue class of m_i in $\mathbb{Z}_n$. Since m_i and n are relatively prime, there holds $m_i^* \in \mathbb{Z}_n^*$. We put $G = \{m_1^*, m_2^*, \ldots, m_s^*\} \subseteq \mathbb{Z}_n^*$. Consider the mapping

$$\alpha: G \longrightarrow Aut_K E.$$
$$m_i^* \to \alpha_{m_i}$$

As $\alpha_{m_i} = \alpha_{m_j} \iff m_i^* = m_j^*$, the mapping α is well defined and one-to-one. Both G and $Aut_K E$ have s elements, so α is also onto $Aut_K E$. Then α has an inverse β:

$$\beta: Aut_K E \longrightarrow G \subseteq \mathbb{Z}_n^*.$$
$$\alpha_{m_i} \to m_i^*$$

Suppose $\alpha_{m_i}\alpha_{m_j} = \alpha_{m_k}$. Then

$$\zeta^{m_k} = \zeta\alpha_{m_k} = \zeta\alpha_{m_i}\alpha_{m_j} = (\zeta\alpha_{m_i})\alpha_{m_j} = (\zeta^{m_i})\alpha_{m_j} = (\zeta\alpha_{m_j})^{m_i} = (\zeta^{m_j})^{m_i} = \zeta^{m_i m_j},$$

so $m_k \equiv m_i m_j \pmod{n}$, so $m_i^* m_j^* = m_k^*$ and therefore

$$(\alpha_{m_i}\alpha_{m_j})\beta = (\alpha_{m_k})\beta = m_k^* = m_i^* m_j^* = (\alpha_{m_i})\beta(\alpha_{m_j})\beta.$$

Hence $\beta: Aut_K E \to \mathbb{Z}_n^\times$ is a one-to-one group homomorphism, and $Im\ \beta = G$ is a subgroup of $\mathbb{Z}_n^\times$ and β is an isomorphism form $Aut_K E$ onto G. This proves that $Aut_K E$ is isomorphic to a subgroup of $\mathbb{Z}_n^\times$. It follows from Lagrange's theorem that $|Aut_K E| = |G|$ divides $|\mathbb{Z}_n^\times| = \varphi(n)$.

(4) Since $Aut_K E$ is isomorphic to a subgroup of $\mathbb{Z}_n^\times$ and $|\mathbb{Z}_n^\times| = \varphi(n)$ is finite, we have the equivalence $Aut_K E \cong \mathbb{Z}_n^\times \iff |Aut_K E| = \varphi(n)$.

We have $|Aut_K E| = deg\ f(x)$ and $\varphi(n) = deg\ \Phi_n(x)$. Now $f(x)$ divides $\Phi_n(x)$ in $K[x]$ and both $f(x)$ and $\Phi_n(x)$ are monic, so $f(x) = \Phi_n(x)$ if and only if $deg\ f(x) = deg\ \Phi_n(x)$, so if and only if $|Aut_K E| = \varphi(n)$.

Finally, since $\Phi_n(x)$ is monic and ζ is a root of $\Phi_n(x)$, irreducibility of $\Phi_n(x)$ in $K[x]$ implies that $\Phi_n(x)$ is the minimal polynomial of ζ over K, i.e., that $f(x) = \Phi_n(x)$. Conversely, if $f(x) = \Phi_n(x)$, then $\Phi_n(x)$ is irreducible. □

When the base field is $\mathbb{Q}$, we have sharper results.

58.11 Theorem: *For any $n \in \mathbb{N}$, the n-th cyclotomic polynomial $\Phi_n(x)$ over $\mathbb{Q}$ is irreducible in $\mathbb{Z}[x]$.*

Proof: Let $n \in \mathbb{N}$ and let $g(x)$ be an irreducible divisor of $\Phi_n(x)$ in $\mathbb{Z}[x]$, with $deg\ g(x) \geqslant 1$ so that $\Phi_n(x) = g(x)h(x)$, say, where $g(x), h(x) \in \mathbb{Z}[x]$ are monic polynomials. Let ζ be a root of $g(x)$. Thus $g(x)$ is the minimal polynomial of ζ over $\mathbb{Q}$.

Our first step will be to show that ζ^p is also a root of $g(x)$ for any prime number p relatively prime to n. Now ζ is a root of $\Phi_n(x)$, so $o(\zeta) = \varphi(n)$ and if p is a prime number such that $(p,n) = 1$, then $o(\zeta^p) = \varphi(n)$ and ζ^p is also a primitive n-th root of unity: ζ^p is a root of $\Phi_n(x)$, so ζ^p is a root of $g(x)$ or of $h(x)$. Let us assume, by way of contradiction, that ζ^p is not a root of $g(x)$. Then ζ^p is a root of $h(x)$. Then ζ is a root of $h(x^p)$ and $h(x^p)$ is divisible by the minimal polynomial $g(x)$ of ζ over $\mathbb{Q}$.

Let us write $h(x^p) = g(x)p(x)$, where $p(x) \in \mathbb{Q}[x]$. Let

$$h(x^p) = g(x)q(x) + r(x), \qquad r(x) = 0 \text{ or } deg\ r(x) < deg\ g(x)$$

be the division algorithm in $\mathbb{Z}[x]$ ($g(x)$ is monic). The uniqueness of the quatient and remainder in $\mathbb{Z}[x] \subseteq \mathbb{Q}[x]$ implies $p(x) = q(x)$ and $r(x) = 0$. Thus we have $h(x^p) = g(x)p(x)$, where $p(x) \in \mathbb{Z}[x]$.

Let $\nu\colon \mathbb{Z} \to \mathbb{F}_p$ be the natural homomorphism and let $\hat{\nu}\colon \mathbb{Z}[x] \to \mathbb{F}_p[x]$ be the homomorphism of Lemma 33.7. We shall write $\overline{s}(x)$ instead of $(s(x))\hat{\nu}$ for $s(x) \in \mathbb{Z}[x]$. Then $h(x^p) = g(x)p(x)$ implies

$$\overline{h}(x^p) = \overline{g}(x)\overline{p}(x) \qquad \text{in } \mathbb{F}_p[x].$$

Since $char\ \mathbb{Z}_p = p$, there holds $\overline{h}(x^p) = \overline{h}(x)^p$ in $\mathbb{F}_p[x]$ and we get

$$\overline{h}(x)^p = \overline{g}(x)\overline{p}(x) \qquad \text{in } \mathbb{F}_p[x].$$

So there is an irreducible factor of $\overline{g}(x)$ in $\mathbb{F}_p[x]$ which divides $\overline{h}(x)^p$ and which therefore divides $\overline{h}(x)$ in $\mathbb{F}_p[x]$. Thus $\overline{g}(x)$ and $\overline{h}(x)$ have a common factor in $\mathbb{F}_p[x]$. Since $g(x)h(x) = \Phi_n(x)$ divides $x^n - 1$ in $\mathbb{Z}[x]$, there is a $k(x)$ in $\mathbb{Z}[x]$ such that

$$g(x)h(x)k(x) = x^n - 1 \qquad \text{in } \mathbb{Z}[x],$$

so

$$\overline{g}(x)\overline{h}(x)\overline{k}(x) = \overline{x^n - 1} = x^n - 1 \qquad \text{in } \mathbb{F}_p[x]$$

and $x^n - 1 \in \mathbb{F}_p[x]$ has a multiple root. But the derivative of $x^n - 1 \in \mathbb{F}_p[x]$ is not $0 \in \mathbb{F}_p[x]$, so relatively prime to $x^n - 1$ and $x^n - 1 \in \mathbb{F}_p[x]$ has no multiple roots. This contradiction shows that ζ^p must be a root of $g(x)$.

Hence if p is a prime number,
$(p,n) = 1$,
ζ is a root of $g(x)$, then ζ^p is a root of $g(x)$.

Let m be any natural number satisfying $1 \leqslant m \leqslant n$ and $(n,m) = 1$. Then $m = p_1^{a_1}p_2^{a_2}\ldots p_r^{a_r}$ with suitable prime numbers p_i relatively prime to n. Repeated application of the result we have just proved shows that ζ^m is a root of $g(x)$ when ζ is. This is true for each of the $\varphi(n)$ natural numbers m such that $1 \leqslant m \leqslant n$ and $(n,m) = 1$. Thus $g(x)$ has $\varphi(n)$ (distinct) roots ζ^m and $g(x)$ is divisible by $\prod_{\substack{1 \leqslant m \leqslant n \\ (n,m) = 1}} (x - \zeta^m) = \Phi_n(x)$. Hence $\Phi_n(x) = g(x)$ and $\Phi_n(x)$ is irreducible in $\mathbb{Z}[x]$. □

58.12 Theorem: *Let $n \in \mathbb{N}$ and let ζ be a primitive n-th root of unity in some extension of $\mathbb{Q}$. Then $\mathbb{Q}(\zeta)$ is Galois over $\mathbb{Q}$* and $Aut_{\mathbb{Q}}\mathbb{Q}(\zeta) \cong \mathbb{Z}_n^\times$.

Proof: Since $\Phi_n(x)$ is monic and irreducible in $\mathbb{Z}[x]$, it is irreducible in $\mathbb{Q}[x]$ (Lemma 34.11). The claim follows now from Theorem 58.10. □

We consider the special case of Theorem 58.12 where n is prime. Let p be a prime number. Then the isomorphism $\mathbb{Z}_p^\times = \mathbb{F}_p^\times \cong Aut_{\mathbb{Q}}\mathbb{Q}(\zeta)$ is given, in the notation of the proof of Theorem 58.10, by $m_i^* \to \alpha_{m_i} \in Aut_{\mathbb{Q}}\mathbb{Q}(\zeta)$, where $\alpha_{m_i}: \zeta \to \zeta^{m_i}$. Both $\mathbb{F}_p^\times$ and $Aut_{\mathbb{Q}}\mathbb{Q}(\zeta)$ are cyclic. Let $g \in \mathbb{Z}$ be such that its residue class $g^* \in \mathbb{F}_p^\times$ is a generator of $\mathbb{F}_p^\times$. Then $Aut_{\mathbb{Q}}\mathbb{Q}(\zeta) = <\sigma>$, where $\sigma = \alpha_g$, i.e., σ is the automorphism $\zeta \to \zeta^g$.

Then the p-th primitive roots of unity are

$$\zeta, \zeta^g, \zeta^{g^2}, \zeta^{g^3}, \ldots, \zeta^{g^{p-2}}$$

and we have $\sigma^k: \zeta \to \zeta^{g^k}$. Let us put $\zeta_k = \zeta^{g^k}$. Then $\zeta_{k+(p-1)} = \zeta\sigma^{k+(p-1)} = \zeta\sigma^k = \zeta_k$ so that any index k can be replaced by any j with $k \equiv j \pmod{p-1}$. Now $\zeta_k\sigma = (\zeta^{g^k})\sigma = (\zeta\sigma)^{g^k} = (\zeta^g)^{g^k} = \zeta^{g^{k+1}} = \zeta_{k+1}$ and $\zeta_k\sigma^m = (\zeta^{g^k})\sigma^m = (\zeta\sigma^m)^{g^k} = (\zeta^{g^m})^{g^k} = \zeta^{g^{k+m}} = \zeta_{k+m}$. Thus σ raises the index by 1 and more generally σ^m raises the index by m.

Let us find the intermediate fields of the extension $\mathbb{Q}(\zeta)/\mathbb{Q}$. Since $\mathbb{Q}(\zeta)$ is Galois over $\mathbb{Q}$, and since $Aut_{\mathbb{Q}}\mathbb{Q}(\zeta) = <\sigma>$ is cyclic of order $p-1$, there is one and only one intermediate field for each positive divisor e of $p-1$, namely the one that corresponds to the subgroup $<\sigma^e>$ of $Aut_{\mathbb{Q}}\mathbb{Q}(\zeta)$. Hence this field, say K_e, is the fixed field of σ^e and $|K_e:\mathbb{Q}| = |<\sigma>:<\sigma^e>| = e$. In order to describe K_e explicitly, we note first that

$$\{1,\zeta,\zeta^g,\zeta^{g^2},\zeta^{g^3},\ldots,\zeta^{g^{p-2}}\} = \{1,\zeta_0,\zeta_1,\zeta_2,\zeta_3,\ldots,\zeta_{p-2}\} = \{1,\zeta,\zeta\sigma,\zeta\sigma^2,\zeta\sigma^3,\ldots,\zeta\sigma^{p-2}\}$$

is a $\mathbb{Q}$-basis of $\mathbb{Q}(\zeta)$ since this set is equal to $\{1,\zeta,\zeta^2,\zeta^3,\ldots,\zeta^{p-1}\}$, which is a $\mathbb{Q}$-basis of $\mathbb{Q}(\zeta)$ by Theorem 50.7. So any element u in $\mathbb{Q}(\zeta)$ can be written in the form

$$u = a_0\zeta_0 + a_1\zeta_1 + a_2\zeta_2 + a_3\zeta_3 + \cdots + a_{p-1}\zeta_{p-1}$$

with uniquely determined $a_0, a_1, a_2, \ldots, a_{p-1} \in \mathbb{Q}$. Here

$$\begin{aligned} u\sigma^e &= (a_0\zeta_0 + a_1\zeta_1 + a_2\zeta_2 + a_3\zeta_3 + \cdots + a_{p-2}\zeta_{p-2})\sigma^e \\ &= a_0\zeta_{e+0} + a_1\zeta_{e+1} + a_2\zeta_{e+2} + a_3\zeta_{e+3} + \cdots + a_{p-2}\zeta_{e+(p-2)} \end{aligned}$$

and u is fixed by σ^e, i.e., $u\sigma^e = u$ if and only if

$$\begin{aligned} a_0\zeta_{e+0} + a_1\zeta_{e+1} + a_2\zeta_{e+2} + a_3\zeta_{e+3} + \cdots + a_{p-2}\zeta_{e+(p-2)} \\ = a_{e+0} + a_{e+1}\zeta_{e+1} + a_{e+2}\zeta_{e+2} + a_{e+3}\zeta_{e+3} + \cdots + a_{e+(p-2)}\zeta_{e+(p-2)} \end{aligned}$$

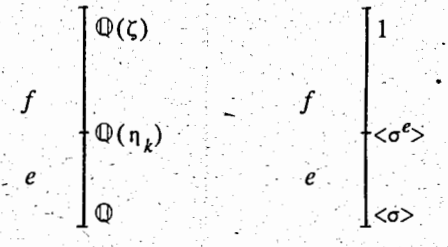
which is equivalent, when we put $f = (p-1)/e$, to

$$\begin{aligned} a_0 &= a_{e+0} = a_{2e+0} = a_{3e+0} = \cdots = a_{(f-1)e+0} \\ a_1 &= a_{e+1} = a_{2e+1} = a_{3e+1} = \cdots = a_{(f-1)e+1} \\ a_2 &= a_{e+2} = a_{2e+2} = a_{3e+2} = \cdots = a_{(f-1)e+2} \\ &\cdots\cdots\cdots\cdots \\ a_{(e-1)} &= a_{e+(e-1)} = a_{2e+(e-1)} = a_{3e+(e-1)} = \cdots = a_{(f-1)e+(e-1)} \end{aligned}$$

and this means $u = a_0(\zeta_0 + \zeta_e + \zeta_{2e} + \zeta_{3e} + \cdots + \zeta_{(f-1)e})$
$+ a_1(\zeta_1 + \zeta_{e+1} + \zeta_{2e+1} + \zeta_{3e+1} + \cdots + \zeta_{(f-1)e+1})$
$+ a_2(\zeta_2 + \zeta_{e+2} + \zeta_{2e+2} + \zeta_{3e+2} + \cdots + \zeta_{(f-1)e+2})$
$+ \cdots$
$+ a_{e-1}(\zeta_{e-1} + \zeta_{e+(e-1)} + \zeta_{2e+(e-1)} + \zeta_{3e+(e-1)} + \cdots + \zeta_{(f-1)e+(e-1)})$.

We put $\eta_k = \zeta_k + \zeta_{e+k} + \zeta_{2e+k} + \zeta_{3e+k} + \cdots + \zeta_{(f-1)e+k}$ $(k = 1,2, \ldots, e-1)$. The elements η_k are called the *periods of f terms*. We see u is fixed by σ^e if and only if $u = a_0\eta_0 + a_1\eta_1 + a_2\eta_2 + \cdots + a_{e-1}\eta_{e-1}$ with $a_0, a_1, a_2, \ldots, a_{e-1} \in \mathbb{Q}$. So $\{\eta_0, \eta_1, \eta_2, \ldots, \eta_{e-1}\}$ is a $\mathbb{Q}$-basis of K_e.

Note that $\sigma: \eta_0 \to \eta_1, \eta_1 \to \eta_2, \eta_2 \to \eta_3, \ldots, \eta_{e-2} \to \eta_{e-1}, \eta_{e-1} \to \eta_0$. Thus each of $\eta_0, \eta_1, \eta_2, \ldots, \eta_{e-1}$ is fixed by σ^e and by powers of σ^e, but not by any other automorphism of $Aut_{\mathbb{Q}}\mathbb{Q}(\zeta)$. Hence all intermediate fields $\mathbb{Q}(\eta_0)$, $\mathbb{Q}(\eta_1)$, $\mathbb{Q}(\eta_2)$, ..., $\mathbb{Q}(\eta_{e-1})$ of $\mathbb{Q}(\zeta)/\mathbb{Q}$ correspond to the same subgroup $\langle\sigma^e\rangle$ of $Aut_{\mathbb{Q}}\mathbb{Q}(\zeta)$. This forces $\mathbb{Q}(\eta_0) = \mathbb{Q}(\eta_1) = \mathbb{Q}(\eta_2) = \cdots = \mathbb{Q}(\eta_{e-1}) = K_e$. So any period of f terms is a primitive element of K_e, the unique intermediate field of $\mathbb{Q}(\zeta)/\mathbb{Q}$ with $|K_e:\mathbb{Q}| = e$.



We summarize our results.

58.13 Theorem: *Let p be a prime number and ζ a primitive p-th root of unity in some extension field of $\mathbb{Q}$. Let $g \in \mathbb{Z}$ be such that its residue class g^* in $\mathbb{F}_p$ is a generator of $\mathbb{F}_p^\times$. Then*

(1) *$\mathbb{Q}(\zeta)$ is Galois over $\mathbb{Q}$;*
(2) *$Aut_{\mathbb{Q}}\mathbb{Q}(\zeta)$ is a cyclic group of order $p-1$. A generator of $Aut_{\mathbb{Q}}\mathbb{Q}(\zeta)$ is the $\mathbb{Q}$-automorphism $\sigma: \zeta \to \zeta^g$.*
(3) *Let e and f be natural numbers such that $ef = p-1$, and put*

$$\eta_k = \zeta^{g^{e+k}} + \zeta^{g^{2e+k}} + \cdots + \zeta^{g^{(f-1)e+k}} \qquad (k = 0,1,2,\ldots,e-1).$$

Then there is one and only one intermediate field of the extension $\mathbb{Q}(\zeta)/\mathbb{Q}$ whose $\mathbb{Q}$-dimension is equal to e. This field is $\mathbb{Q}(\eta_k)$ for any $k = 0,1,2,\ldots,e-1$. The set $\{\eta_0, \eta_1, \eta_2, \ldots, \eta_{e-1}\}$ is a $\mathbb{Q}$-basis of $\mathbb{Q}(\eta_k)$. All intermediate fields of $\mathbb{Q}(\zeta)/\mathbb{Q}$ are found in this way as e ranges through the positive divisors of $p-1$. □

58.14 Examples: (a) We find all intermediate fields of $\mathbb{Q}(\zeta)$, where the complex number $\zeta \in \mathbb{C}$ is a primitive 7-th root of unity. These are the simple extensions of $\mathbb{Q}$ whose primitive elements are the periods. In order to construct the periods, we need a generator of $\mathbb{F}_7^\times$. One checks easily that the residue class of 3 is a generator of $\mathbb{F}_7^\times$. The images of ζ under powers of the automorphism $\sigma: \zeta \to \zeta^3$ are

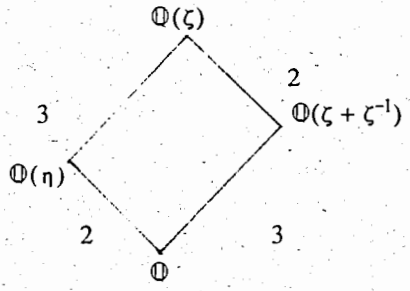
$$\zeta, \zeta^3, \zeta^2, \zeta^6, \zeta^4, \zeta^5.$$

The 1-term periods are $\zeta, \zeta^3, \zeta^2, \zeta^6, \zeta^4, \zeta^5$ and $\mathbb{Q}(\zeta) = \mathbb{Q}(\zeta^3) = \mathbb{Q}(\zeta^2) = \mathbb{Q}(\zeta^6) = \mathbb{Q}(\zeta^4) = \mathbb{Q}(\zeta^5)$ is the intermediate field with $|\mathbb{Q}(\zeta):\mathbb{Q}| = 6$.

The 2-term periods are $\zeta + \zeta^6, \zeta^3 + \zeta^4, \zeta^2 + \zeta^5$ and $\mathbb{Q}(\zeta + \zeta^6)$ is the intermediate field $|\mathbb{Q}(\zeta + \zeta^6):\mathbb{Q}| = 3$. We also have $\mathbb{Q}(\zeta + \zeta^6) = \mathbb{Q}(\zeta + \zeta^{-1}) = \mathbb{Q}(\zeta^3 + \zeta^4) = \mathbb{Q}(\zeta^2 + \zeta^5)$.

The 3-term periods are $\eta = \zeta + \zeta^2 + \zeta^4$, $\eta' = \zeta^3 + \zeta^6 + \zeta^5$ and $\mathbb{Q}(\eta) = \mathbb{Q}(\eta')$ is the intermediate field with $|\mathbb{Q}(\eta):\mathbb{Q}| = 2$.

The 6-term period is $\zeta^3 + \zeta^2 + \zeta^6 + \zeta^4 + \zeta^5 + \zeta = -1$ and $\mathbb{Q}(-1) = \mathbb{Q}$ is the intermediate field with $|\mathbb{Q}(-1):\mathbb{Q}| = 1$.



(b) We determine the intermediate fields of $\mathbb{Q}(\zeta)/\mathbb{Q}$, where ζ is a primitive 17-th root of unity. The divisors of $17 - 1 = 16$ are 1,2,4,8,16 and there are five intermediate fields, of dimensions 1,2,4,8,16 over $\mathbb{Q}$.

The residue class of $3 \in \mathbb{Z}$ in $\mathbb{F}_{17}^{\times}$ is a generator of $\mathbb{F}_{17}^{\times}$. The successive powers of 3 are congruent, modulo 17, to

1,3,9,10,13,5,15,11,16,14,8,7,4,12,2,6

The 8-term periods are

$$\eta_0 = \zeta + \zeta^9 + \zeta^{13} + \zeta^{15} + \zeta^{16} + \zeta^8 + \zeta^4 + \zeta^2$$
$$\eta_1 = \zeta^3 + \zeta^{10} + \zeta^5 + \zeta^{11} + \zeta^{14} + \zeta^7 + \zeta^{12} + \zeta^6.$$

An elementary computation shows that $\eta_0 + \eta_1 = -1$ and $\eta_0\eta_1 = -4$. So η_0 and η_1 are the roots of $x^2 + x - 4$. Hence $\eta_0, \eta_1 = \frac{-1 \mp \sqrt{17}}{2}$. Which of η_0, η_1 has the plus sign depends on the choice of ζ. We may assume ζ is a 17-th root of unity that appears in the period with the plus sign (otherwise replace ζ by one of the roots of unity that appear in the period with the plus sign). Then $\eta_0 = \frac{-1+\sqrt{17}}{2}$ and $\eta_1 = \frac{-1-\sqrt{17}}{2}$.

The 4-term periods are

$$\chi_0 = \zeta + \zeta^{13} + \zeta^{16} + \zeta^4, \quad \chi_2 = \zeta^9 + \zeta^{15} + \zeta^8 + \zeta^2$$

$$\chi_1 = \zeta^3 + \zeta^5 + \zeta^{14} + \zeta^{12}, \quad \chi_3 = \zeta^{10} + \zeta^{11} + \zeta^7 + \zeta^6$$

and $\quad \chi_0 + \chi_2 = \eta_0, \chi_0\chi_2 = -1; \qquad \chi_1 + \chi_3 = \eta_1, \chi_1\chi_3 = -1.$

Hence χ_0 and χ_2 are the roots of $x^2 - \eta_0 x - 1$ and χ_1 and χ_3 are the roots of $x^2 - \eta_1 x - 1$. Here we may put $\chi_0 = \frac{\eta_0+\sqrt{\eta_0^2+4}}{2}$ and $\chi_2 = \frac{\eta_0-\sqrt{\eta_0^2+4}}{2}$ by assuming that ζ is a 17-th root of unity that appears in the period χ with the plus sign The signs of radicals in $\chi_1,\chi_3 = \frac{\eta_1\mp\sqrt{\eta_1^2+4}}{2}$, however, can no longer be arbitrarily assigned by choosing ζ suitably. To determine which of χ_1,χ_3 has the positive radical, we note

$$(\chi_0 - \chi_2)(\chi_1 - \chi_3) = 2(\eta_0 - \eta_1)$$

$$\frac{\sqrt{\eta_0^2+4}}{2} \cdot (\chi_1 - \chi_3) = \sqrt{17},$$

so that $\chi_1 - \chi_3$ is positive. This gives $\chi_1 = \frac{\eta_1+\sqrt{\eta_1^2+4}}{2}$ and $\chi_3 = \frac{\eta_1-\sqrt{\eta_1^2+4}}{2}$.

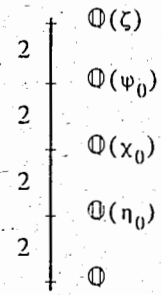
The 2-term periods are

$$\psi_0 = \zeta + \zeta^{16}, \qquad \psi_4 = \zeta^{13} + \zeta^4$$
$$\psi_1 = \zeta^3 + \zeta^{14} \qquad \psi_5 = \zeta^5 + \zeta^{12}$$
$$\psi_2 = \zeta^9 + \zeta^8, \qquad \psi_6 = \zeta^{15} + \zeta^2$$
$$\psi_3 = \zeta^{10} + \zeta^7 \qquad \psi_7 = \zeta^{11} + \zeta^6.$$

Here $\psi_0 + \psi_4 = \chi_0$ and $\psi_0\psi_4 = \chi_1$, so ψ_0 and ψ_4 are roots of $x^2 - \chi_0 x + \chi_1$. Thus $\psi_0, \psi_4 = \frac{\chi_0\mp\sqrt{\chi_0^2-4\chi_1}}{2}$. We put $\psi_0 = \frac{\chi_0+\sqrt{\chi_0^2-4\chi_1}}{2}$. In like manner as above, one can find polynomials whose roots are ψ_j and determine the roots without ambiguity.

A 1-term period is ζ, which is a root of $x^2 - \psi_0 x + 1$. Hence we may put $\zeta = \frac{\psi_0+\sqrt{\psi_0^2-4}}{2}$.

The subfield structure of $\mathbb{Q}(\zeta)$ is depicted below.



*

* *

We now prove an important theorem due to J. H. M. Wedderburn which states that any finite division ring is commutative. The proof makes use of the class equation (Lemma 25.16) of the multiplicative group of nonzero elements in a finite division ring. Let us recall the class equation of any finite group G is

$$|G| = \sum_{i=1}^{k} |G:C_G(x_i)|,$$

where k is the number of distinct conjugacy classes, $x_1, x_2, \ldots, x_k$ are representatives of these classes and $C_G(x_i) = \{g \in G: x_i g = g x_i\}$ are the centralizers of x_i $(i = 1,2, \ldots, k)$.

In addition to these centralizer groups, we consider centralizer rings and evaluate their dimensions to find the terms in the class equation. An argument involving cyclotomic polynomials shows than that the class equation cannot hold unless the division ring is commutative.

In order not to interrupt the main argument, we establish two lemmas we will need.

58.15 Lemma: *Let n be a natural number greater than one and let $\Phi_n(x)$ be the n-th cyclotomic polynomial over $\mathbb{Q}$. Then, for any a proper divisor d of n, we have*

$$\Phi_n(x) \left| \frac{x^n - 1}{x^d - 1} = x^{(n/d)-1} + x^{(n/d)-2} + \cdots + x^{(n/d)} + 1 \right. \qquad \text{in } \mathbb{Z}[x]$$

and, for any natural number q,

$$\Phi_n(q) \left| \frac{q^n - 1}{q^d - 1} \right. \qquad \text{in } \mathbb{Z}.$$

Proof: Since $\Phi_n(x) | (x^n - 1)$ and $x^n - 1 = (x^d - 1)[(x^n - 1)/(x^d - 1)]$, it is sufficient to show that $\Phi_n(x)$ is relatively prime to $x^d - 1$ for any proper divisor d of n. But this is clear, because $\Phi_n(x)$ and $x^d - 1$ have no root in common: the roots of $\Phi_n(x)$ are primitive n-th roots of unity, whereas a root of $x^d - 1$ cannot be a primitive n-th root of unity if d is a *proper* divisor of n. This proves the divisibility relation in $\mathbb{Z}[x]$. Substituting any integer q for x (and using $\Phi_n(x)$, $(x^n - 1)/(x^d - 1) \in \mathbb{Z}[x]$) we obtain the divisibility relation in $\mathbb{Z}$. □

58.16 Lemma: *If $n > 1$ and $\Phi_n(x)$ is the n-th cyclotomic polynomial over $\mathbb{Q}$, then $|\Phi_n(q)| > q - 1$ for all $q \in \mathbb{N}$ with $q \geqslant 2$.*

Proof: We have $\Phi_n(x) = \prod_{\substack{k=1 \\ (k,n)=1}}^{n} (x - \zeta^k)$, where ζ is a primitive n-th root of unity in some extension field of $\mathbb{Q}$. For example, we may take $\zeta = e^{2\pi i/n}$. Substituting q for x and using the triangle inequality $|a - b| \geqslant \big||a| - |b|\big|$, we get

$$|\Phi_n(q)| = \prod_{\substack{k=1 \\ (k,n)=1}}^{n} |q - \zeta^k| = \prod_{\substack{k=1 \\ (k,n)=1}}^{n} |q - e^{2\pi k i/n}| \geqslant \prod_{\substack{k=1 \\ (k,n)=1}}^{n} \big||q| - |e^{2\pi k i/n}|\big|$$

$$= \prod_{\substack{k=1 \\ (k,n)=1}}^{n} |q - 1| = (q-1)^{\varphi(n)} = (q-1)\cdot(q-1)^{\varphi(n)-1} > q - 1$$

in case $\varphi(n) - 1 \geqslant 1$ since $q > 1$. In case $\varphi(n) - 1 = 0$, we have $n = 2$ and $|\Phi_2(q)| = q + 1 > q - 1$. □

58.17 Theorem (Wedderburn's theorem): *If D is a finite division ring, then D is a field.*

Proof: Let D be a division ring with finitely many elements. $D^\times = D \setminus \{0\}$ is then a finite group under multiplication and the class equation of $D^\times$ is

$$|D^\times| = \sum_{i=1}^{k} |D^\times : C_{D^\times}(x_i)|,$$

where k is the number of distinct conjugacy classes of $D^\times$ and $x_1, x_2, \ldots, x_k$ are representatives of these classes.

We now put $C_D(x_i) = \{a \in D : x_i a = a x_i\} = C_{D^\times}(x_i) \cup \{0\} \subseteq D$. Since $a, b \in C_D(x_i)$ implies $x_i(a + b) = x_i a + x_i b = a x_i + b x_i = (a + b)x_i$, we see $C_D(x_i)$ is closed under addition and thus $C_D(x_i)$ is a subgroup of D under addition (Lemma 9.3(2)). As $C_D(x_i)\setminus\{0\} = C_{D^\times}(x_i)$ is a subgroup of $D^\times$, we conclude that $C_D(x_i)$ is a division ring (a subdivision ring of D).

The same argument proves that the *center* of the *ring* D:

$$Z = \{a \in D : xa = ax \text{ for all } x \in D\} = Z(D^\times) \cup \{0\}$$

is a a subdivision ring of D. But Z is a commutative division ring, i.e., Z is a field. Then *char* $Z = p$ for some prime number p and $|Z| = p^t$ for some natural number t. We put $q = p^t = |Z|$ for brevity.

We have $Z \subseteq C_D(x_i) \subseteq D$. Since multiplication in D is associative and distributive over addition, and since $1a = a$ for all $a \in C_D(x_i)$, we get that $C_D(x_i)$ and D are vector spaces over Z. Let $dim_Z C_D(x_i) = m_i$ and $dim_Z D = n$. Then, as in Lemma 52.1, we have $|C_D(x_i)| = |Z|^{m_i} = q^{m_i}$ and $|D| = |Z|^n = q^n$. This gives $|C_{D^\times}(x_i)| = |C_D(x_i)\setminus\{0\}| = |C_D(x_i)| - 1 = q^{m_i} - 1$ and likewise $|D^\times| = |D\setminus\{0\}| = |D| - 1 = q^n - 1$. The class equation is therefore

$$q^n - 1 = \sum_{i=1}^{k} |D^\times : C_{D^\times}(x_i)| = \sum_{i=1}^{k} \frac{|D^\times|}{|C_{D^\times}(x_i)|} = \sum_{i=1}^{k} \frac{q^n - 1}{q^{m_i} - 1}.$$

Now $|D^*:C_{D^*}(x_i)|$ is an integer, so $q^{m_i} - 1$ divides $q^n - 1$ and this implies that m_i divides n (Lemma 52.7(1)).

We want to show that D is commutative, or, what is the same thing, that $Z = D$. We will assume $Z \neq D$ and derive a contradiction. Well, if $Z \neq D$, then $n > 1$ and there is at least one x_i such that $|D^*:C_{D^*}(x_i)| \neq 1$, because $|D^*:C_{D^*}(x_i)| = 1$ if and only if $x_i \in Z(D^*)$. We so choose the notation that $\{x_1, x_2, \ldots, x_h\} = Z(D^*)$ and $x_{h+1}, \ldots, x_k$ are not in the center of D^*. Then the class equation becomes

$$q^n - 1 = \sum_{i=1}^{h} |D^*:C_{D^*}(x_i)| + \sum_{i=h+1}^{k} |D^*:C_{D^*}(x_i)| = |Z(D^*)| + \sum_{i=h+1}^{k} \frac{q^n - 1}{q^{m_i} - 1}$$

$$q^n - 1 = (q - 1) + \sum_{i=h+1}^{k} \frac{q^n - 1}{q^{m_i} - 1}$$

and m_i is a *proper* divisor of n for $i = h + 1, \ldots, k$. As $n > 1$ by assumption, $\Phi_n(q)$ divides $\frac{q^n - 1}{q^{m_i} - 1}$ for all $i = h + 1, \ldots, k$ (Lemma 58.15); $\Phi_n(q)$ divides also $q^n - 1$. We read from the class equation that $\Phi_n(q)$ divides $q - 1$. But this is impossible, for $|\Phi_n(q)| > q - 1$ by Lemma 58.16.

Thus $n = 1$ and $D = Z$ is commutative. □

Exercises

1. Find the m-th cyclotomic polynomial $\Phi_m(x)$ over $\mathbb{Q}$ for $m \leqslant 50$.

2. Let $\Phi_m(x)$ denote the m-th cyclotomic polynomial over $\mathbb{Q}$. Prove:
   (a) $\Phi_{2n}(x) = \Phi_n(-x)$ if $2 \nmid n$.
   (b) $\Phi_{pn}(x) = \Phi_n(x^p)/\Phi_n(x)$ if p is an odd prime number and $p \nmid n$.

3. Evaluate the p^k-th cyclotomic polynomial $\Phi_{p^k}(x)$ over $\mathbb{Q}$ if p is a prime number and $k \in \mathbb{N}$.

4. Let $p, k \in \mathbb{N}$ and p be prime. Let $\Phi_p(x)$ denote the p-th cyclotomic polynomial over $\mathbb{Q}$. Prove that, if $d | \Phi_p(k)$, then $d \equiv 1 \pmod{p}$ or $d = p$.

5. Let $p \in \mathbb{N}$ be prime, $k \in \mathbb{Z}$ and let k^* be the residue class of k in $\mathbb{F}_p$. Let $n \in \mathbb{N}$ and $\Phi_n(x)$ the n-th cyclotomic polynomial over $\mathbb{Q}$. Suppose that $p \nmid n$. Prove the following statements.

(a) $p \mid \Phi_n(k)$ if and only if $o(k^*) = n$ (order of k^* in $\mathbb{F}_p^\times$ is n).

(b) There is an integer a with $p \mid \Phi_n(a)$ if and only if $p \equiv 1 \pmod{n}$.

6. Let $n \in \mathbb{N}$ and $\Phi_n(x)$ the n-th cyclotomic polynomial over $\mathbb{Q}$ and let $p_1, p_2, \ldots, p_m$ be prime numbers of the form $tn + 1$ $(t, n \in \mathbb{Z})$. Use Ex. 5 and prove the following statements.

(a) $\Phi_n(anp_1p_2 \ldots p_m) \equiv \mp 1 \pmod{np_1p_2 \ldots p_m}$ for any $a \in \mathbb{N}$.

(b) $\Phi_n(anp_1p_2 \ldots p_m) \neq \mp 1$ if $a \in \mathbb{N}$ is sufficiently large.

(c) For some $a \in \mathbb{N}$, there is a prime divisor p of $\Phi_n(anp_1p_2 \ldots p_m)$ which is distinct from $p_1, p_2, \ldots, p_m$.

(d) There are infinitely many prime numbers p of the form $tn + 1$.

(This is a special case of the following celebrated theorem of Dirichlet: if a, b are any relatively prime integers, then there infinitely many prime numbers of the form $an + b$.)

7. Find all subfields of $\mathbb{Q}(\zeta)$, where $\zeta \in \mathbb{C}$ is a primitive n-th root of unity and $n = 4,5,6,8,12$. Prove $e^{2\pi i/5} = \dfrac{-1+\sqrt{5}+i\sqrt{10+2\sqrt{5}}}{4}$.

8. Prove the formula due to Gauss:

$$\cos\frac{2\pi}{17} = -\frac{1}{16} + \frac{1}{16}\sqrt{17} + \frac{1}{16}\sqrt{34 - 2\sqrt{17}} + \frac{1}{8}\sqrt{17 + 3\sqrt{17} - \sqrt{34 - 2\sqrt{17}} - 2\sqrt{34 + 2\sqrt{17}}}$$

9. Under the hypotheses of Theorem 58.13, show that the *set* of periods independent of the integer g for which g^* is a generator of $\mathbb{F}_p^\times$, but the indices of *individual* periods do depend on g. Describe this dependence.

10. Let the hypotheses of Theorem 58.13 be valid, with p an odd prime number, and let η_0, η_1 be the $[(p-1)/2]$-term periods. Prove that $\eta_0\eta_1 = -(p-1)/4$ or $(p+1)/4$ according as $p \equiv 1 \pmod{p}$ or $p \equiv 3 \pmod{p}$. Show that $\eta_0 - \eta_1 = \mp\sqrt{(-1)^{(p-1)/2}p}$. (The sign depends on the primitive p-th root of unity ζ we take. If we choose $\zeta = e^{2\pi i/p} \in \mathbb{C}$, then the sign is plus. This is considerably difficult to prove. This exercise shows $\mathbb{Q}(\mp\sqrt{(-1)^{(p-1)/2}p})$ is contained in the cyclotomic field $\mathbb{Q}(\zeta)$. A theorem of class field theory,

known as Kronecker-Weber theorem, states that any finite dimensional Galois extension of $\mathbb{Q}$ whose Galois group is abelian is contained in a suitable cyclotomic extension of $\mathbb{Q}$.)

11. Let $\zeta_k \in \mathbb{C}$ denote a primitive k-th root of unity. Show that, if $(n,m) = 1$, then $\mathbb{Q}(\zeta_n,\zeta_m) = \mathbb{Q}(\zeta_{nm})$ and $\mathbb{Q}(\zeta_n) \cap \mathbb{Q}(\zeta_m) = \mathbb{Q}$.

12. Let $\zeta \in \mathbb{C}$ be a primitive n-th root of unity. Prove that all roots of unity in $\mathbb{Q}(\zeta)$ are $\pm\zeta^j$ $(j = 0,1,2, \ldots,n-1)$.

13. Let $p \in \mathbb{N}$ be a prime number and $\Phi_p(x)$ the p-th cyclotomic polynomial over $\mathbb{Q}$. Find the discriminant of $\Phi_p(x)$.

14. Show that any finite subring of a division ring is a division ring.

§ 59
Applications

This paragraph consists of five parts. In the first part, we give an exact definition of solvability by radicals, discuss radical extensions and establish the fundamental theorem due to Galois that a polynomial equation is solvable by radicals if and only if the Galois group of the polynomial is a solvable group. In the second part, we apply this theorem to the general polynomial of degree n over a field and deduce Abel's theorem: if $n \geqslant 5$, then the general polynomial of degree n is not solvable by radicals. In the third part, we discuss solvability of equations when the degree is prime. In the fourth part, we give formulas for the roots of polynomials of degree two, three and four. In the last part, we examine which real numbers can be constructed by ruler and compass.

*

* *

We study solvability of polynomials by an algebraic formula. We start by clarifying what we mean by an algebraic formula. Intuitively, this is an expression like

$$\sqrt[u]{\sqrt[k]{\sqrt[m]{\ldots}+\sqrt[n]{\ldots}}+\sqrt[l]{\sqrt[r]{\ldots}-\sqrt[s]{\ldots}}}$$

involving addition, subtraction, multiplication, division and taking n-th roots, where the terms in innermost radicals are elements of the field to which the coefficients of the polynomial belong. If the terms are from a field K, the field operations addition, subtraction, multiplication, division give rise to elements in the same field K, but extraction of n-th root $\sqrt[n]{a}$ amounts to a field extension, namely to the adjunction of a root of $x^n - a$ to K. Thus a formula basically desribes a sequence

$$K_0 \subseteq K_1 \subseteq K_2 \subseteq \ldots \subseteq K_n$$

of fields, K_0 being the field in which the coefficients of the given polynomial lie and each K_{i+1} is obtained from K_i by adjoining a root of a polynomial of the form $x^n - a \in K_i[x]$ to K_i. These considerations lead to the following definitions.

59.1 Definition: Let E/K be a field extension. If there are elements $u_1, u_2, \ldots, u_n$ in E such that

(1) $E = K(u_1, u_2, \ldots, u_n)$,

(2) there exist natural numbers $h_1, h_2, \ldots, h_m$ such that $u_1^{h_1} \in K$ and $u_i^{h_i} \in K(u_1, \ldots, u_{i-1})$ for $i = 2, \ldots, n$,

then E is called a *radical extension of* K.

59.2 Definition: Let K be a field and $f(x) \in K[x]$. We say *the equation* $f(x) = 0$ *is solvable by radicals* provided there is a splitting field S of $f(x)$ over K and a radical extension R of K such that $K \subseteq S \subseteq R$.

Note we do not require the splitting field S itself to be a radical extension, rather that S be contained in some radical extension.

It follows from Definition 59.1 that a radical extension is a finitely generated and in fact a finite dimensional extension. When we consider radical extensions as in Definition 59.1 we agree, for uniformity in notation, to read $K(u_1, \ldots, u_{h-1})$ as K when $h = 1$.

If, in the setup of Definition 59.1, $h_i = rs$ and if we put $u_i^r = u_i$, so that $u_i^s \in K(u_1, \ldots, u_{i-1})$, then we may insert the field $K(u_1, \ldots, u_{i-1}, u_i)$ between $K(u_1, \ldots, u_{i-1})$ and $K(u_1, \ldots, u_{i-1}, u_i)$:

$$K(u_1, \ldots, u_{i-1}) \subseteq K(u_1, \ldots, u_{i-1}, u_i) \subseteq K(u_1, \ldots, u_{i-1}, u_i, u_i) = K(u_1, \ldots, u_{i-1}, u_i),$$

without disturbing the condition (2) in Definition 59.1 because

$$u_i^s \in K(u_1, \ldots, u_{i-1}) \text{ and } u_i^r \in K(u_1, \ldots, u_{i-1}, u_i).$$

Thus inserting additional intermediate fields if necessary, we may suppose that the h_i in Definition 59.1 are prime numbers whenever we want to.

One of the principle theorems in this paragraph is that, if a polynomial equation $f(x) = 0$ is solvable by radicals, then the Galois group of $f(x)$ is a solvable group (Definition 27.19). In fact, we obtain more general results. In the next three lemmas, we study radicality of some related field extensions.

59.3 Lemma: *Let $K \subseteq L \subseteq E$ be fields.*
(1) *If E is a radical extension of K, then E is a radical extension of L.*
(2) *If L is a radical extension of K and if E is a radical extension of L, then E is a radical extension of K.*

Proof: (1) If E is a radical extension of K, there are $u_1, u_2, \ldots, u_n$ in E such that $E = K(u_1, u_2, \ldots, u_n)$ and $u_i^{h_i} \in K(u_1, \ldots, u_{i-1})$ for some natural numbers h_i $(i = 1, 2, \ldots, n)$. Then $E = L(u_1, u_2, \ldots, u_n)$, as $K \subseteq L \subseteq E$ and also $u_1^{h_1} \in L$ and $u_i^{h_i} \in L(u_1, \ldots, u_{i-1})$ for $i = 2, \ldots, n$. Thus E is a radical extension of L.

(2) If L is a radical extension of K, then there are elements $u_1, u_2, \ldots, u_n$ in L such that $L = K(u_1, u_2, \ldots, u_n)$ and natural numbers $h_1, h_2, \ldots, h_m$ such that $u_1^{h_1} \in K$ and $u_i^{h_i} \in K(u_1, \ldots, u_{i-1})$. If E is a radical extension of L, then there are elements $t_1, t_2, \ldots, t_m$ in E such that $E = L(t_1, t_2, \ldots, t_n)$ and natural numbers $k_1, k_2, \ldots, k_m$ such that $t_1^{k_1} \in L$ and $t_i^{k_i} \in L(t_1, \ldots, t_{i-1})$. Thus there are elements $u_1, u_2, \ldots, u_n, t_1, t_2, \ldots, t_m$ in E such that
$E = K(u_1, u_2, \ldots, u_n, t_1, t_2, \ldots, t_m)$ and natural numbers $h_1, h_2, \ldots, h_m, k_1, k_2, \ldots, k_m$ such that

$$u_1^{h_1} \in K \text{ and}$$
$$u_i^{h_i} \in K(u_1, \ldots, u_{i-1}) \text{ for } i = 2, \ldots, n,$$
$$t_1^{k_1} \in K(u_1, u_2 \ldots, u_{n-1}, u_n)$$
$$t_i^{k_i} \in K(u_1, u_2, \ldots, u_n, t_1, \ldots, t_{i-1}) \text{ for } i = 2, \ldots, m.$$

This shows that E is a radical extension of K. □

59.4 Lemma: *Let K be a field and L, M radical extensions of K contained in some extension of K. Then their compositum* (see Definition 50.17) *LM is a radical extension of K.*

Proof: Since L and M are radical extensions of K, we have

$L = K(u_1,u_2, \ldots,u_n)$ and $u_i^{h_i} \in K(u_1, \ldots,u_{h-1})$ $\quad (i = 1,2, \ldots,n)$

$M = K(t_1,t_2, \ldots,t_m)$ and $t_j^{k_j} \in K(t_1, \ldots,t_{k-1})$ $\quad (j = 1,2, \ldots,m)$

with some suitable elements u_i, t_j and natural numbers h_i,k_j. Now LM is the smallest subfield of E containing K and $u_1,u_2, \ldots,u_n,t_1,t_2, \ldots,t_m$, so $LM = K(u_1,u_2, \ldots,u_n,t_1,t_2, \ldots,t_m)$. Since $u_i^{h_i} \in K(u_1, \ldots,u_{h-1})$ for $i = 1,2, \ldots,n$ and likewise $t_j^{k_j} \in K(u_1,u_2, \ldots,u_n,t_1, \ldots,t_{k-1})$ for $j = 1,2, \ldots,m$, we conclude that LM is a radical extension of K (where $K(u_1,u_2, \ldots,u_n,t_1, \ldots,t_{k-1})$ is to be read as $K(u_1,u_2, \ldots,u_n)$ when $k = 1$). □

59.5 Lemma: *Let E/K be a radical field extension and let N be a normal closure of K over E. Then N is a radical extension of K.*

Proof: Let $\{a_1,a_2, \ldots,a_m\}$ be a K-basis of E and let $f_i(x) \in K[x]$ be the minimal polynomial of a_i over K. We remind the reader of the fact that N is a splitting field of $f(x) := f_1(x)f_2(x) \ldots f_m(x)$ over K (see the proof of Theorem 55.11).

Let a be a root of $f_j(x)$. There is a K-isomorphism $\varphi: K(a_j) \to K(a)$ with $a_j\varphi = a$ (Theorem 53.2). Since N is a splitting field of $f(x)$ over $K(a_j)$ and over $K(a)$ (Example 53.5(e)), the isomorphism φ extends to a K-automorphism $\Phi: N \to N$ of N (Theorem 53.7). Then $E\Phi$ is an intermediate field of N/K which is K-isomorphic to E and $E\Phi$ contains the root a_j of $f_j(x)$. In this way, we find, for each $j = 1,2, \ldots,m$ and for each root b of $f_j(x)$, intermediate fields of N/K which are K-isomorphic to E and which contain the root b of $f_j(x)$.

Let $E_1,E_2, \ldots,E_s$ be the fields obtained in this way. Then each E_i is K-isomorphic to E and so a radical extension of K. Using Lemma 59.4 repeatedly, we get that the compositum $E_1(E_2(E_3(\ldots E_s))\ldots)$ is a radical extension of K. But this compositum is a subfield of N containing all roots of $f(x)$. Since N is a splitting field of $f(x)$ over K, the compositum must equal N. Thus N is a radical extension of K. □

59.6 Lemma: *Let E/K be a finite dimensional field extension, $m \in \mathbb{N}$. Assume char $K = 0$ or $(m, \text{char } K) = 1$ and let ζ be a primitive m-th root of unity. If E is Galois over K, then $E(\zeta)$ is also Galois over K.*

Proof: We have a chain of fields $K \subseteq E \subseteq E(\zeta)$. By Theorem 55.7, there is a polynomial $f(x) \in K[x]$ whose irreducible factors are separable over K such that E is a splitting field of $f(x)$ over K and $E(\zeta)$ is the splitting field of the m-th cyclotomic polynomial $\Phi_m(x)$ over E, whose irreducible factors, too, are separable over E (Theorem 58.10).

We claim $E(\zeta)$ is a splitting field of $f(x)\Phi_m(x) \in K[x]$ over K (we have $\Phi_m(x) \in K[x]$ by Lemma 58.7(2)). Since the irreducible factors of $f(x)\Phi_m(x)$ have no multiple roots, they are separable over K and the claim will imply that $E(\zeta)$ is a Galois extension of K (Theorem 55.7).

Any root of $f(x)\Phi_m(x)$ is in $E(\zeta)$, so $f(x)\Phi_m(x)$ splits in $E(\zeta)$. Now let F be a subfield of $E(\zeta)$ containing K such that $f(x)\Phi_m(x)$ splits in F. Then all roots of $f(x)$ are in F and, since E is generated over K by the roots of $f(x)$ (Example 53.5(d)), $E \subseteq F$. Moreover, F contains ζ, so we have $E(\zeta) \subseteq F$. Thus $f(x)\Phi_m(x)$ cannot split in any proper subfield of $E(\zeta)$ containing K and $E(\zeta)$ is therefore Galois over K. □

59.7 Lemma: *Let K be a field, $n \in \mathbb{N}$ and assume that char $K = 0$ or (char K, n) = 1. Suppose that K contains a primitive n-th root of unity. Let $a \in K\setminus\{0\}$ and let u be a root of $x^n - a \in K[x]$. Then*

(1) $K(u)$ is a cyclic extension of K;
(2) $|K(u):K|$ divides n and $u^{|K(u):K|} \in K$.

Proof: (1) We must show $K(u)$ is Galois over K and $Aut_K K(u)$ is a cyclic group. If $\zeta \in K$ is a primitive n-th root of unity, then $u, \zeta u, \zeta^2 u, \ldots, \zeta^{n-1}u$ are the roots of $x^n - a$. Thus $K(u)$ is a splitting field of $x^n - a$ over K. The polynomial $x^n - a$ has no multiple roots, so the irreducible divisors of $x^n - a$ are separable over K. Thus $K(u)$ is Galois over K (Theorem 55.7).

We now show that $Aut_K K(u)$ is cyclic. If $\sigma \in Aut_K K(u)$, then $u\sigma$ is a root of $x^n - a$, so $u\sigma = \zeta_\sigma u$ for some (not necessarily primitive) n-th root ζ_σ of unity. Since $\zeta_{\sigma\tau}u = u(\sigma\tau) = (u\sigma)\tau = (\zeta_\sigma u)\tau = (\zeta_\sigma \tau)(u\tau) = \zeta_\sigma \cdot \zeta_\tau u = (\zeta_\sigma \zeta_\tau)u$ and so

$\zeta_{\sigma\tau} = \zeta_\sigma \zeta_\tau$ for any $\sigma, \tau \in Aut_K K(u)$, the mapping $\varphi: Aut_K K(u) \to \langle\zeta\rangle \subseteq K^*$ is a homomorphism of groups. Here $\sigma \in Ker\,\varphi$ if and only if $\zeta_\sigma = 1$, i.e., if and only if $u\sigma = u$, so if and only if σ is the identity mapping on $K(u)$. Thus $Ker\,\varphi = 1$ and φ is one-to-one. This shows that $Aut_K K(u)$ is isomorphic to a subgroup of K^*. Since $Aut_K K(u)$ is finite, $Aut_K K(u)$ is a cyclic group by Theorem 52.18.

$\sigma \longrightarrow \zeta_\sigma$

(2) Let $|K(u):K| = d$. Since $K(u)$ is Galois over K, we have $|Aut_K K(u)| = d$ by the fundamental theorem of Galois theory. So $Aut_K K(u)$ is a cyclic group of order d, say $Aut_K K(u) = \langle\sigma\rangle$. Now $\langle\sigma\rangle$ is isomorphic to a subgroup $\langle\zeta_\sigma\rangle$ of $\langle\zeta\rangle$ and $\langle\zeta\rangle$ has order n. Hence $d|n$. Moreover, $o(\zeta_\sigma) = |\langle\zeta_\sigma\rangle| = |\langle\sigma\rangle| = o(\sigma) = d$, so $\zeta_\sigma{}^d = 1$ and $(u^d)\sigma = (u\sigma)^d = (\zeta_\sigma u)^d = \zeta_\sigma{}^d u^d = u^d$, so u^d is fixed by σ and by $Aut_K K(u)$, so $u^d \in K$ since $K(u)$ is Galois over K. □

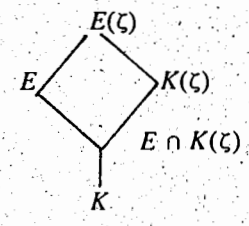
We now proceed to prove that the Galois group of a polynomial is a solvable group if the polynomial is solvable by radicals. It will be seen that it is sufficient to prove this under the asumption that a splitting field of the polynomial is a radical extension (rather than a subfield of a radical extension), and one may moreover suppose that splitting field of the polynomial is Galois over the base field. As a technical convenience, we will bring a certain root of unity into the base field. Then the subgroups of the Galois group corresponding to the intermediate fields as in Definition 59.1 under the Galois correspondence will make up a chain such that each group will be normal in the next one and the factor groups will be cyclic by Lemma 59.7. This will give an abelian series of the Galois group, which must be therefore solvable.

59.8 Theorem: *Let K be a field and E a Galois extension of K. If E is a radical extension of K, then $Aut_K E$ is a solvable group.*

Proof: Since E is a radical extension of K, we have $E = K(u_1, u_2, \ldots, u_n)$ and there are natural numbers $h_1, h_2, \ldots, h_m$ such that $u_i^{h_i} \in K(u_1, \ldots, u_{i-1})$ for $i = 1, 2, \ldots, n$. Without loss of generality, we may suppose h_i are prime numbers.

First we show that *char* K, if distinct from 0, can be assumed to be distinct from all the prime numbers h_i. Indeed, if $0 \neq char\ K = p = h_i$, then $u_i^p \in K(u_1, \ldots, u_{i-1})$. But E is Galois, hence separable over K and over $K(u_1, \ldots, u_{i-1})$ (Lemma 55.6), so u_i is separable over $K(u_1, \ldots, u_{i-1})$ and $K(u_1, \ldots, u_{i-1}, u_i) = K(u_1, \ldots, u_{i-1})(u_i) = K(u_1, \ldots, u_{i-1})(u_i^p) = K(u_1, \ldots, u_{i-1}, u_i^p) = K(u_1, \ldots, u_{i-1})$ by Lemma 55.16. Thus $K(u_1, \ldots, u_{i-1}) = K(u_1, \ldots, u_{i-1}, u_i)$ and u_i can be deleted from the set of generators. We assume all generators of this type have been deleted and thus all the prime numbers h_i are relatively prime to the characteristic of K in case $char\ K = p \neq 0$.

Put $m = h_1 h_2 \ldots h_n$ and let ζ be a primitive m-th root of unity. We consider the cyclotomic extensions $E(\zeta)$ of E and $K(\zeta)$ of K:



Since either *char* $K = 0$ or *char* K is relatively prime to m, Lemma 59.6 shows that $E(\zeta)/K$ is Galois (E is finite dimensional over K because E is a radical extension of K). Theorem 54.25(2) gives: $Aut_E E(\zeta) \trianglelefteq Aut_K E(\zeta)$ and $Aut_K E \cong Aut_K E(\zeta) / Aut_E E(\zeta)$. We want to prove that $Aut_K E$ is a solvable group. If we can show that $Aut_K E(\zeta)$ is solvable, then $Aut_K E$ will also be solvable, because a factor group of a solvable group is solvable (Lemma 27.20). Thus it is sufficient to prove that $Aut_K E(\zeta)$ is a solvable group.

We make one further reduction. $K(\zeta)$ is a Galois extension of K by Theorem 58.10(2), so $Aut_{K(\zeta)} E(\zeta) \trianglelefteq Aut_K E(\zeta)$ and moreover $Aut_K K(\zeta) \cong Aut_K E(\zeta) / Aut_{K(\zeta)} E(\zeta)$. We know that $Aut_K K(\zeta)$ is abelian (Theorem 58.10(3)). Thus $Aut_K E(\zeta) / Aut_{K(\zeta)} E(\zeta)$ is abelian and solvable. If we can show that $Aut_{K(\zeta)} E(\zeta)$ is solvable, then $Aut_K E(\zeta)$ will also be solvable in view of Lemma 27.21. Thus it is sufficient to prove that $Aut_{K(\zeta)} E(\zeta)$ is a solvable group.

We put $K(\zeta) = E_0$ and $K(\zeta, u_1, \ldots, u_i) = E_i$ for $i = 1, 2, \ldots, n$. In particular $E_n = K(\zeta, u_1, u_2, \ldots, u_n) = K(u_1, u_2, \ldots, u_n)(\zeta) = E(\zeta)$. Since E_n/K is Galois, E_n is Galois over any intermediate field (Theorem 54.25(1)). Thus E_n is Galois over E_0.

Let $G_i \leqslant Aut_{E_0}E_n = Aut_{K(\zeta)}E(\zeta)$ be the subgroup $E_i' = Aut_{E_i}E_n$ of $Aut_{E_0}E_n$ corresponding to $E_i\ (i = 0,1,2,\ldots,n)$:

$$\begin{array}{rl|l} E(\zeta) = E_n & & G_n = 1 \\ & & \\ E_i & & G_i \\ & & \\ E_{i-1} & & G_{i-1} \\ & & \\ K(\zeta) = E_0 & & G_0 \end{array}$$

Now $\quad char\ E_{i-1} = 0$ or relatively prime to h_i

$$E_i = K(\zeta,u_1,\ldots,u_i) = K(\zeta,u_1,\ldots,u_{i-1})(u_i) = E_{i-1}(u_i),$$

$$u_i^{h_i} \in K(u_1,\ldots,u_{i-1}) \subseteq K(\zeta,u_1,\ldots,u_{i-1}) = E_{i-1},$$

and $\quad E_{i-1}$ has a primitive h_i-th root of unity,

since in fact E_{i-1} has a primitive m-th root of unity $(i = 1,2,\ldots,n)$. Thus Lemma 59.7 applies and shows that E_i is a cyclic extension of E_{i-1} of degree $|E_i : E_{i-1}| = h_i$ or 1. In particular, E_i is Galois over E_{i-1} and, since E_n is also Galois over E_{i-1}, we get $G_i \trianglelefteq G_{i-1}$ and $G_{i-1}/G_i \cong Aut_{E_{i-1}}E_i$ from Theorem 54.25(2). Thus $|G_{i-1}/G_i| = |E_i : E_{i-1}| = h_i$ or 1 and G_{i-1}/G_i is cyclic (of prime order h_i or of order 1). Hence

$$1 = G_n \trianglelefteq G_{n-1} \trianglelefteq G_{n-2} \trianglelefteq \ldots \trianglelefteq G_1 \trianglelefteq G_0 = Aut_{E_0}E_n = Aut_{K(\zeta)}E(\zeta)$$

is an abelian series of $Aut_{K(\zeta)}E(\zeta)$ and $Aut_{K(\zeta)}E(\zeta)$ is a solvable group. This completes the proof. □

59.9 Lemma: *Let E/K be a field extension and*

$$K_1 = \{a \in S : a\varphi = a \text{ for all } \varphi \in Aut_K E\}.$$

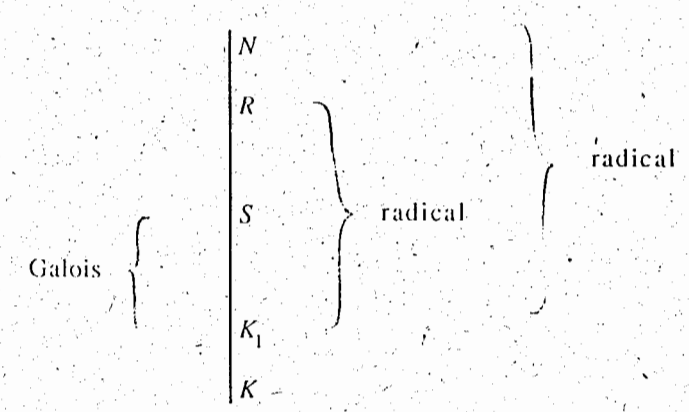
Then $Aut_K E = Aut_{K_1} E$ and E is Galois over K_1.

Proof: Clearly K_1 is closed under addition, subtraction, multiplication and division; so K_1 is a field and we have $K \subseteq K_1$ by the very definition of K_1. Any K-automorphism of E fixes the elements of K_1 and, since $K \subseteq K_1$, any K_1-automorphism of E fixes K elementwise. Thus $Aut_K E = Aut_{K_1} E$.

If $b \in E$ is fixed by all K_1-automorphisms of E, then b is fixed by all K-automorphisms of E, so $b \in K_1$. Hence K_1 is the fixed field of $Aut_{K_1}E$, which means E is a Galois extension of K_1. □

59.10 Theorem: *Let $K \subseteq S \subseteq R$ be a fields. If R is a radical extension of K, then $Aut_K S$ is a solvable group.*

Proof: We put $K_1 = \{a \in S : a\varphi = a \text{ for all } \varphi \in Aut_K E\}$. Then $Aut_K S = Aut_{K_1} S$ and S is a Galois extension of K_1 by Lemma 59.9. Moreover, R is a radical extension of K_1 (Lemma 59.3(1)). Let N be a normal closure of K_1 over R. Then N is a radical extension of K_1 by Lemma 59.5.



Now S is a Galois extension of K_1, so S is (K_1, N)-stable (Theorem 54.23). Then $Aut_S N \trianglelefteq Aut_{K_1} N$ and $Aut_{K_1} N / Aut_S N$ is isomorphic to the subgroup of $Aut_K S$ consisting of all K_1-automorphisms of S that are extendible to N (Theorem 54.24). What is this subgroup of $Aut_{K_1} S$? Since N is normal over K_1, there is a polynomial $f(x)$ in $K_1[x]$ such that N is a splitting field of $f(x)$ over K_1 (Theorem 55.8). Thus N is a splitting field of $f(x)$ over S (Example 53.5(e)) and any K_1-automorphism of S can be extended to a K-automorphism of N (Theorem 53.7). So the subgroup of $Aut_{K_1} S$ consisting of all K_1-automorphisms of S that are extendible to N is actually the whole $Aut_K S$. We get

$$Aut_K S = Aut_{K_1} S \cong Aut_{K_1} N \,/\, Aut_S N.$$

Since any factor group of a solvable group is solvable (Lemma 27.20), it suffices to prove that $Aut_{K_1} N$ is solvable. As in the first paragraph in this proof, we replace the base field by another and make the extension Galois. We put $K_2 = \{a \in N : a\varphi = a \text{ for all } \varphi \in Aut_{K_1} N\} \subseteq N$. Then $K_1 \subseteq K_2$ and $Aut_{K_1} N = Aut_{K_2} N$. Here N is a Galois, radical extension of K_2 (Lemma 59.9, Lemma 59.3(1)) and Theorem 59.8 yields that $Aut_{K_1} N = Aut_{K_2} N$ is a solvable group. □

From Definition 59.2 and Theorem 59.10, we get

59.11 Theorem: *Let K be a field and $f(x) \in K[x]$. If the equation $f(x) = 0$ is solvable by radicals, then the Galois group of $f(x)$ is a solvable group.* □

We now want to establish the converse of Theorem 59.9. Let E/K be a Galois extension. If $Aut_K E$ is solvable, then the composition factors of $Aut_K E$ are cyclic of prime order and the Galois correspondence gives rise to a chain of intermediate fields in which the two consecutive terms represent a cyclic extension of prime degree. There are two types of cyclic extensions of prime degree: (1) extensions of the form $K(u)/K$ where u is a root of $x^p - x$ and $char\, K = 0$ or $(p, char\, K) = 1$ and (2) extensions of the form $K(u)/K$, where u is a root of $x^p - x - a$ and $p = char\, K$. (Just as Lemma 59.7 is the converse of Theorem 57.11, Theorem 57.10 admits a converse; see § 57, Ex. 3,4.) Hence the extensions of the second type will creep into the intermediate field structure of E/K. There are two ways of coping with this situation. Either we modify the definition of radical extensions so as to include extensions of the second type as admissible intermediate steps (see Ex. 1,2.) or we impose restrictive hypotheses on the characteristic to prevent extensions of the second type from coming up.

59.12 Theorem: *Let K be a field and E a finite dimensional Galois extension of K. Assume $char\, K = 0$ or $0 \neq char\, K$ and $char\, K$ does not divide $|E:K|$. If $Aut_K E$ is a solvable group, then there is a radical extension R of K such that $K \subseteq E \subseteq R$.*

Proof: We make induction on $|E:K|$. If $|E:K| = 1$, then $E = K$ and K is a radical extension of K containing K. Thus the theorem is proved in case $|E:K| = 1$.

Let $n = |E:K|$. Assume $n \geqslant 2$ and the theorem is proved for all field extensions of degree $< n$.

Since $Aut_K E$ is a solvable group, $Aut_K E =: G$ has a subgroup of prime index, say $H \leqslant G$ and $|G:H| = p$, where p is a prime number (Theorem 27.25). Here p divides $|Aut_K E| = |E:K|$, so $p \neq char\, K$ by hypothesis. Let ζ be a primitive p-th root of unity. The cyclotomic extension $K(\zeta)$ is a radical extension of K, so, if we can prove there is a radical extension R of $K(\zeta)$

$$
\left|\begin{array}{lll}
1 & & E(\zeta) \\
 & E & K(\zeta) \\
 & & E \cap K(\zeta) \\
H & & \\
p & & \\
Aut_K E & & K
\end{array}\right.
$$

containing $E(\zeta)$, then R will be a radical extension of K containing E (Lemma 59.3(2)).

We show that $E(\zeta)$ is contained in some radical extension of $K(\zeta)$. Since E is Galois over K, Lemma 59.6 yields $E(\zeta)$ is Galois over K and Theorem 54.23 yields E is $(K, E(\zeta))$-stable. So the restriction mapping $\sigma \to \sigma_{|E}$ is a homomorphism $\varphi: Aut_{K(\zeta)} E(\zeta) \to Aut_K E$. If $\sigma \in Ker\, \varphi \subseteq Aut_{K(\zeta)} E(\zeta)$, then $\sigma_{|E}$ fixes all elements of E, so σ fixes all elements of E and also ζ, so σ is the identity mapping on $E(\zeta)$, so $Ker\, \varphi = \{\iota_{E(\zeta)}\}$ and φ is one-to-one.

We distinguish two cases, according as $Im\, \varphi$ is a proper subgroup of $Aut_K E$ or equal to $Aut_K E$. Since $E(\zeta)$ is Galois over K, it is Galois over $K(\zeta)$ by Theorem 54.25(1) and so $|E(\zeta):K(\zeta)| = |Aut_{K(\zeta)} E(\zeta)|$.

If $Im\ \varphi < Aut_K E$, then $|E(\zeta):K(\zeta)| = |Aut_{K(\zeta)}E(\zeta)| = |Im\ \varphi| < |Aut_K E| = n$. Now $Aut_{K(\zeta)}E(\zeta)$, being isomorphic to a subgroup of the solvable group $Aut_K E$, itself is a solvable group (Lemma 27.20) and $E(\zeta)$ is Galois over $K(\zeta)$, so, by induction, there is a radical extension R of $K(\zeta)$ containing $E(\zeta)$. The proof is complete in this case.

If $Im\ \varphi = Aut_K E$, then φ is an isomorphism and has an inverse isomorphism $\varphi^{-1}: Aut_K E \to Aut_{K(\zeta)}E(\zeta)$. We put $J = H\varphi^{-1}$. Then $J \trianglelefteq Aut_{K(\zeta)}E(\zeta)$ and $|Aut_{K(\zeta)}E(\zeta):J| = p$. Since H is solvable, its isomorphic image J is solvable. Let $F = (Aut_J E(\zeta))'$ be the intermediate field of the Galois extension $E(\zeta)/K(\zeta)$ corresponding to J.

$$\begin{array}{ccccc} & E(\zeta) & \longleftrightarrow & 1 & \overset{\varphi}{\to} & 1 \\ & & & & & \\ & F & \longleftrightarrow & J & & H \\ p & & & p & p & \\ & K(\zeta) & \longleftrightarrow & Aut_{K(\zeta)}E(\zeta) & & Aut_K E \end{array}$$

As $J \trianglelefteq Aut_{K(\zeta)}E(\zeta)$, Theorem 54.25(2) shows that F is Galois over K and $Aut_{K(\zeta)}F \cong Aut_{K(\zeta)}E(\zeta) / Aut_F E(\zeta) = Aut_{K(\zeta)}E(\zeta) / J \cong C_p$. So F is a cyclic extension of $K(\zeta)$, so $F = K(u)$ for some root of a suitable polynomial of the form $x^p - a$ in $K(\zeta)[x]$ (Theorem 57.11). Thus F is a radical extension of $K(\zeta)$. Here $Aut_F E(\zeta) = J$ is solvable, $E(\zeta)$ is Galois over F (Theorem 54.25(1)) and $|E(\zeta):F| < |E(\zeta):F||F:K(\zeta)| = |E(\zeta):K(\zeta)| = n$, so, by induction, there is a radical extension R of F with $F \subseteq E(\zeta) \subseteq R$. Since R is a radical extension of F and F is a radical extension of $K(\zeta)$, Lemma 59.3(2) yields that R is a radical extension of $K(\zeta)$ with $K(\zeta) \subseteq E(\zeta) \subseteq R$. This completes the proof. □

59.13 Theorem: *Let K be a field and $f(x) \in K[x]$ a polynomial of degree $n > 0$. Suppose $charK = 0$ or $0 \neq charK > n$. Then the equation $f(x) = 0$ is solvable by radicals if and only if the Galois group of $f(x)$ is a solvable group.*

Proof: If the equation $f(x) = 0$ is solvable by radicals, then the Galois group of the polynomial $f(x)$ is solvable by Theorem 59.11.

Conversely, let S be a splitting field of $f(x)$ over K and assume that the Galois group $Aut_K S$ of $f(x)$ is solvable. In order to prove that the equation $f(x) = 0$ is solvable by radicals, i.e., in order to prove that there is a radical extension R of K satisfying $K \subseteq S \subseteq R$, it suffices, in view of Theorem 59.12, to show that S is Galois over K and $char\, K = 0$ or $char\, K$ does not divide $|S:K|$.

To prove that S is Galois over K, we use Theorem 55.7. We need only show that the irreducible factors of $f(x)$ are separable over K. This is clear in case $char\, K = 0$. If $char\, K \neq 0$ and $ax^d + \cdots \in K[x]$ is an irreducible factor of $f(x)$ with $a \neq 0$, then $d \leq n < char\, K$ and $da \neq 0 \in K$, so its derivative $dax^{d-1} + \cdots$ is not equal to $0 \in K[x]$ and $ax^d + \cdots$ is separable over K.

Thus we are done in case $char\, K = 0$. In case $char\, K \neq 0$, we have $char\, K > n$, so the prime number $char\, K$ does not divide $n!$ and, as $|S:K| \leq n!$, it does not divide $|S:K|$ either. The proof is complete. □

*

* *

In this part, we prove the celebrated theorem due to Abel which states that the general polynomial (over a field of characteristic 0) of degree n is solvable by radicals if and only if $n \leq 4$ and some related results. First of all, we must explain what we mean by the general polynomial of degree n.

59. 14 Definition: Let K be a field and let $a_1, a_2, \ldots, a_{n-1}, a_n$ be n distinct indeterminates over K. The polynomial

$$g(x) = x^n - a_1 x^{n-1} + a_2 x^{n-2} - a_3 x^{n-3} + - \cdots + (-1)^{n-1} a_{n-1} x + (-1)^n a_n$$

in $K(a_1, a_2, \ldots, a_{n-1}, a_n)[x]$ is called the *general polynomial of degree n over K*.

Any monic polynomial in $K[x]$ can be obtained from $f(x)$ by substituting appropriate elements of K for the indeterminates. This justifies the

terminology. Note, however, a peculiarity: the general polynomial of degree n over K is *not* a polynomial over K, that is, it is not in $K[x]$, but in $K(a_1,a_2,\ldots,a_{n-1},a_n)[x]$.

Alternating signs are attached to the coefficients a_j for convenience in computations. This makes it easier to compare the coefficients a_j with the elementary symmetric polynomials.

Our main goal is to prove that the Galois group of the general polynomial of degree n is the symmetric group S_n. After we established some prepatory lemmas, we prove that each permutation of the roots induces an automorphism of the splitting field if the roots are indeterminates (Theorem 59.17) and that we can indeed treat the roots of the general polynomial as indeterminates (Theorem 59.18).

59.15 Lemma: *Let D_1,D_2 be integral domains and F_1,F_2 the field of fractions of D_1,D_2, respectively. If $\varphi: D_1 \to D_2$ is a ring isomorphism, then the mapping*

$$\begin{aligned} \varphi_1: F_1 &\longrightarrow F_2 \\ a/b &\to (a\varphi)/(b\varphi) \end{aligned}$$

is a field isomorphism.

Proof: We are to show that φ_1 is a one-to-one ring homomorphism from F_1 onto F_2. Let $a,b,c,d \in D_1$ and $b,d \neq 0$. Then

$$\begin{aligned} a/b = c/d &\iff ad = bc \iff (ad)\varphi = (bc)\varphi \iff a\varphi\cdot d\varphi = b\varphi\cdot c\varphi \\ &\iff (a\varphi)/(b\varphi) = (c\varphi)/(d\varphi) \iff (a/b)\varphi_1 = (c/d)\varphi_1, \end{aligned}$$

which shows that φ_1 is well defined and one-to-one. Moreover, if $u \in F_2$, then $u = e/f$ for some $e,f \in D_2$ with $f \neq 0$, then $e = a\varphi$ and $f = b\varphi$ for some $a,b \in D_1$ and $b \neq 0$, so $u = e/f = a\varphi/b\varphi = (a/b)\varphi_1$ is the image of $a/b \in F_1$ under φ_1 and thus φ_1 is onto F_2.

It remains to prove that φ_1 preserves addition and multiplication. This is easy: if $a,b,c,d \in D$ and $b,d \neq 0$, then

$$\begin{aligned} [(a/b) + (c/d)]\varphi_1 &= [(ad + bc)/bd]\varphi_1 = (ad + bc)\varphi/(bd)\varphi \\ &= (a\varphi d\varphi + b\varphi c\varphi)/(b\varphi d\varphi) = (a\varphi/b\varphi) + (c\varphi/d\varphi) \end{aligned}$$

$$= (a/b)\varphi_1 + (c/d)\varphi_1$$

and

$$[(a/b)(c/d)]\varphi_1 = (ac/bd)\varphi_1 = (ac)\varphi/(bd)\varphi = (a\varphi c\varphi)/(b\varphi d\varphi)$$
$$= (a\varphi/b\varphi)(c\varphi/d\varphi) = (a/b)\varphi_1(c/d)\varphi_1.$$

Thus φ_1 is a field isomorphism. □

59.16 Lemma: *Let K be a field and let $x_1, x_2, \ldots, x_n$ be n distinct indeterminates over K.*

(1) *For each permutation $\sigma \in S_n$, the mapping*

$$\sigma' : K(x_1, x_2, \ldots, x_n) \longrightarrow K(x_1, x_2, \ldots, x_n)$$
$$f(x_1, x_2, \ldots, x_n)/g(x_1, x_2, \ldots, x_n) \to f(x_{1\sigma}, x_{2\sigma}, \ldots, x_{n\sigma})/g(x_{1\sigma}, x_{2\sigma}, \ldots, x_{n\sigma})$$

is a field automorphism of $K(x_1, x_2, \ldots, x_n)$.

(2) *If $\sigma, \tau \in S_n$ and $\sigma \neq \tau$, then $\sigma' \neq \tau'$.*

Proof: (1) Let $\sigma \in S_n$. The mapping $\sigma'' : K[x_1, x_2, \ldots, x_n] \to K[x_1, x_2, \ldots, x_n]$
$$f(x_1, x_2, \ldots, x_n) \to f(x_{1\sigma}, x_{2\sigma}, \ldots, x_{n\sigma})$$

is the substitution homomorphism that substitutes $x_{j\sigma}$ for x_j $(j = 1,2,\ldots,n)$. It has an inverse $(\sigma'')^{-1} = (\sigma^{-1})'' : f(x_1, x_2, \ldots, x_n) \to f(x_{1\sigma^{-1}}, x_{2\sigma^{-1}}, \ldots, x_{n\sigma^{-1}})$. Thus σ'' is a ring isomorphism from the integral domain $K[x_1, x_2, \ldots, x_n]$ onto itself. Lemma 59.15 gives that

$$(\sigma'')_1 : K(x_1, x_2, \ldots, x_n) \longrightarrow K(x_1, x_2, \ldots, x_n)$$
$$f(x_1, x_2, \ldots, x_n)/g(x_1, x_2, \ldots, x_n) \mapsto f(x_1, x_2, \ldots, x_n)\sigma''/g(x_1, x_2, \ldots, x_n)\sigma''$$

is a field automorphism of the the field $K(x_1, x_2, \ldots, x_n)$ of fractions of $K[x_1, x_2, \ldots, x_n]$. But $(\sigma'')_1$ is nothing else than σ'. Hence σ' is a field automorphism of $K(x_1, x_2, \ldots, x_n)$.

(2) If $\sigma \neq \tau$, then there is a $j \in \{1,2, \ldots, n\}$ such that $j\sigma \neq j\tau$, then $x_j\sigma' = x_{j\sigma} \neq x_{j\tau} = x_j\tau$, so $\sigma' \neq \tau'$. □

59.17 Theorem: *Let K be a field and $x_1, x_2, \ldots, x_n$ be n distinct indeterminates over K and let*

$$f_1 = \sum x_i$$

$$f_2 = \sum x_i x_j$$

$$f_3 = \sum x_i x_j x_k$$

$$\cdots\cdots\cdots\cdots$$

$$f_n = x_1 x_2 \ldots x_n$$

be the elementary symmetric polynomials in $K[x_1, x_2, \ldots, x_n]$. *Then the field of rational functions* $K(x_1, x_2, \ldots, x_n)$ *is a Galois extension of* $K(f_1, f_2, \ldots f_n)$, *the subfield of* $K(x_1, x_2, \ldots, x_n)$ *generated by* $f_1, f_2, \ldots f_n$ *over* K *and* $Aut_{K(f_1, f_2, \ldots f_n)} K(x_1, x_2, \ldots, x_n) \cong S_n$.

Proof: We put $E = K(x_1, x_2, \ldots, x_n)$ and $L = K(f_1, f_2, \ldots f_n)$. Let x be a new indeterminate over K. If

$$h(x) = x^n - f_1(x_1, x_2, \ldots, x_n)x^{n-1} + f_2(x_1, x_2, \ldots, x_n)x^{n-2} - + \cdots + (-1)^n f_n(x_1, x_2, \ldots, x_n),$$

then $h(x) \in L[x]$ and $h(x)$ splits in E:

$$h(x) = (x - x_1)(x - x_2)\ldots(x - x_n).$$

Since $E = L(x_1, x_2, \ldots, x_n)$ is generated by the roots $x_1, x_2, \ldots, x_n$ of $g(x)$ over L, we deduce that E is a splitting field of $h(x)$ over L (Example 53.5(d)). As $h(x)$ has no multiple roots, the irreducible factors of $h(x)$ in $L[x]$ are separable over L. Theorem 55.7 tells now E is a Galois extension of L.

For each of the $n!$ permutations σ in S_n, there is a $\sigma' \in Aut(E)$ by Lemma 59.16, and σ' fixes $f_1, f_2, \ldots f_n$ as $f_1, f_2, \ldots f_n$ are symmetric polynomials, so σ' fixes $L = K(f_1, f_2, \ldots f_n)$. This means $\sigma' \in Aut_L E$. As $\sigma', \tau' \in Aut_L E$ are distinct whenever $\sigma, \tau \in S_n$ are distinct, there are at least $n!$ automorphisms in $Aut_L E$ and $|Aut_L E| \geqslant n!$. On the other hand, $|Aut_L E| = |E:L|$ since E is Galois over L and $|E:L| \leqslant n!$ by Theorem 53.6 and Theorem 53.8. So we have $|Aut_L E| = n!$. We know from Theorem 56.14 that $Aut_L E$ is isomorphic to a subgroup of S_n. In view of $|Aut_L E| = n!$, it must be isomorphic to S_n. □

59.18 Theorem: *Let K be a field and $n \in \mathbb{N}$. The Galois group of the general polynomial of degree n over K is isomorphic to S_n.*

Proof: Let $a_1, a_2, \ldots, a_{n-1}, a_n$ indeterminates over K so that

$$g(x) = x^n - a_1x^{n-1} + a_2x^{n-2} - a_3x^{n-3} + - \cdots + (-1)^{n-1}a_{n-1}x + (-1)^n a_n$$

is the general polynomial of degree n over K. We put $L_1 = K(a_1, a_2, \ldots, a_n)$. Let E_1 be a splitting field of $f(x)$ over $L_1[x]$ and $r_1, r_2, \ldots, r_n \in E_1[x]$ the roots of $f(x)$. Then $E_1 = L_1(r_1, r_2, \ldots, r_n) = K(a_1, a_2, \ldots, a_n, r_1, r_2, \ldots, r_n)$ $= K(r_1, r_2, \ldots, r_n)$ by Example 53.5(d). The Galois group of $f(x)$ is $Aut_{L_1}E_1$.

Let $x_1, x_2, \ldots, x_n$ be n indeterminates over K which are distinct from the $a_1, a_2, \ldots, a_n$. Let $E = K(x_1, x_2, \ldots, x_n)$ and let $f_1, f_2, \ldots f_n$ be the elementary symmetric polynomials in $K[x_1, x_2, \ldots, x_n]$ and put $L = K(f_1, f_2, \ldots f_n)$. We know $Aut_L E \cong S_n$ from Theorem 59.17.

$$\begin{array}{c|c|c} K(r_1, r_2, \ldots, r_n) = E_1 & & K(x_1, x_2, \ldots, x_n) = E \\ | & & | \\ K(a_1, a_2, \ldots, a_n) = L_1 & \varphi_1 & K(f_1, f_2, \ldots f_n) = L \\ | & & | \\ K & & K \end{array}$$

We show that there is a K-isomorphism $\varphi_1: L_1 \to L$. First observe that we have the substitution homomorphism $\varphi: K[a_1, a_2, \ldots, a_n] \to K[f_1, f_2, \ldots f_n]$ that maps a_i to f_i and $h(a_1, a_2, \ldots, a_n)$ to $h(f_1, f_2, \ldots f_n)$. Clearly φ fixes all elements of K. Furthermore, φ is one-to-one, for if $h_1, h_2 \in K[a_1, a_2, \ldots, a_n]$, then $h_1 \neq h_2$ implies $h_1\varphi = h_1(f_1, f_2, \ldots f_n) \neq h_2(f_1, f_2, \ldots f_n) = h_2\varphi$ by the uniqueness assertion in the fundamental theorem on symmetric polynomials (Theorem 38.4). Thus φ is a ring isomorphism from $K[a_1, a_2, \ldots, a_n]$ onto $Im\ \varphi$. Using Lemma 59.15, we extend φ to a field isomorphism φ_1 from $L = K(a_1, a_2, \ldots, a_n)$ onto the the field of fractions of $Im\ \varphi \subseteq L$. Since $L = K(f_1, f_2, \ldots f_n)$ and $\{f_1, f_2, \ldots f_n\} \subseteq Im\ \varphi$, it follows that the field of fractions of $Im\ \varphi$ is equal to L and thus φ_1 is onto L. Also φ_1 fixes every element of K. Hence $\varphi_1: L_1 \to L$ is a K-isomorphism.

The homomorphism $\hat{\varphi}_1 : L_1[x] \to L[x]$ of Lemma 33.7 maps

$$g(x) = x^n - a_1x^{n-1} + a_2x^{n-2} - a_3x^{n-3} + - \cdots + (-1)^{n-1}a_{n-1}x + (-1)^n a_n$$

to $$h(x) = x^n - f_1x^{n-1} + f_2x^{n-2} - f_3x^{n-3} + - \cdots + (-1)^{n-1}f_{n-1}x + (-1)^n f_n.$$

Here E_1 is a splitting field of $g(x)$ over L_1 and $E = L(x_1, x_2, \ldots, x_n)$ is a splitting field of $h(x)$ over L (Example 53.5(d)), so the isomorphism $\varphi_1 : L_1 \to L$ can be extended to an isomorphism $\varphi : E_1 \to E$ (Theorem 53.7). Lemma 56.11(1) and Theorem 59.17 give now $Aut_{L_1}E_1 \cong Aut_L E \cong S_n$. This completes the proof. □

59.19 Theorem (Abel): *Let K be a field, $n \in \mathbb{N}$ and $g(x)$ the general polynomial of degree n over K. If the equation $g(x) = 0$ is solvable by radicals, then $n \leqslant 4$. Conversely, if $char\, K = 0$ and $n \leqslant 4$, then the equation $g(x) = 0$ is solvable by radicals.*

Proof: The Galois group of $g(x)$ is S_n (Theorem 59.18). If the equation $g(x) = 0$ is solvable by radicals, then S_n is a solvable group (Theorem 59.11), so $n \leqslant 4$ by Theorem 27.26. Conversely, if $n \leqslant 4$ and $char\, K = 0$, then S_n is a solvable group (Example 27.10(a),(b), Theorem 27.25) and the equation $g(x) = 0$ is solvable by radicals (Theorem 59.13). □

Theorem 59.19 is a statement about general polynomials. It does not state that specific polynomial equations of degree $\geqslant 5$ cannot be solvable by radicals.

*

* *

In this part, we examine solvability by radicals of polynomial equations of prime degree. It is necessary to understand the solvable transitive subgroups of S_p. These have a simple structure. After we gave a characterization of solvable transitive subgroups of S_p, we prove the curious result of Galois: "In order for an irreducible equation of prime degree to be solvable by radicals, it is necessary and sufficient that once

any two of the roos are known the others can be deduced from them rationally." (Edwards' translation.)

Let p be a prime number. It will be convenient to regard S_p as acting on the p elements $1,2, \ldots, p$ of $\mathbb{F}_p$. For any $a \in \mathbb{F}_p^*$ and $b \in \mathbb{F}_p$, We write

$$\begin{aligned} \sigma_{a,b}: \mathbb{F}_p &\to \mathbb{F}_p: \\ u &\to au + b \end{aligned}$$

Clearly $\sigma_{a,b} \neq \sigma_{c,d}$ whenever $(a,b) \neq (c,d)$. For any $(a,b),(c,d) \in \mathbb{F}_p^* \times \mathbb{F}_p$, we have

$$\begin{aligned} u\sigma_{a,b}\sigma_{c,d} = (au + b)\sigma_{c,d} &= c(au + b) + d = cau + eb + d \\ &= (ac)u + (bc + d) = u\sigma_{ac,bc+d}, \end{aligned}$$

so $\sigma_{a,b}\sigma_{c,d} = \sigma_{ac,bc+d}$. So $A(p) := \{\sigma_{a,b} : a \in \mathbb{F}_p^*, b \in \mathbb{F}_p\}$ is closed under the composition of mappings. Observe that $\sigma_{1,0} \in A(p)$ is the identity mapping and hence $\sigma_{(1/a),(-b/a)} \in A(p)$ is the inverse of $\sigma_{a,b}$. As the composition of mappings is associative, $A(p)$ is a group. In particular, each $\sigma_{a,b}$ is one-to-one and onto, and can be considered as a permutation in S_p. Thus we shall regard $A(p)$ as a subgroup of S_p. Then

$$\sigma_{a,b} = \begin{pmatrix} 1 & 2 & \ldots & p \\ a+b & a2+b & \ldots & ap+b \end{pmatrix}$$

where the integers ought to be interpreted modulo p. The permutation $\sigma_{0,1} = (12\ldots p)$ of order p will be denoted as π.

59.20 Definition: Let p be a prime number and $\sigma \in S_p$. If there are elements $a \in \mathbb{F}_p^*$ and $b \in \mathbb{F}_p$ such that

$$\sigma = \begin{pmatrix} 1 & 2 & \ldots & p \\ a+b & a2+b & \ldots & ap+b \end{pmatrix}$$

then σ is called a *linear permutation in* S_p. In this case, we shall denote the permutation σ as $\sigma_{a,b}$ Then $A(p) = \{\sigma_{a,b} : a \in \mathbb{F}_p^*, b \in \mathbb{F}_p\}$ is a subgroup of S_p and is called the *one dimensional affine group over* $\mathbb{F}_p$. If $\pi = (12\ldots p)$ and $\langle\pi\rangle \leqslant G \leqslant A(p)$, then G is called a *linear subgroup of* S_p.

59.21 Lemma: *Let p be a prime number, $\pi = (12\ldots p) \in S_p$ and let H be a subgroup of S_p.*

(1) If H is a linear subgroup of S_p, then the only elements of order p in H are $\pi, \pi^2, \pi^3, \ldots, \pi^{p-1}$.

(2) If $\pi, \pi^2, \pi^3, \ldots, \pi^{p-1}$ are the only elements of order p in H, then $\langle\pi\rangle$ is a characteristic and normal subgroup of H.

(3) If $\langle\pi\rangle$ is a normal subgroup of H, then H is a linear subgroup of S_p.

(4) If H is a linear subgroup of S_p and $H \trianglelefteq K \leqslant S_p$, then K is a linear subgroup of S_p.

Proof: (1) Assume H is a linear subgroup of S_p and let $\sigma \in H$. Then $\sigma = \sigma_{a,b}$ for suitable $a,b \in \mathbb{F}_p$, $a \neq 0$.

If $a = 1$, then $u\sigma = u + b = u\pi^b$ for any $u \in \{1,2, \ldots, p-1\}$, so $\sigma = \pi^b$ and $o(\sigma) = o(\pi^b)$ and $o(\pi^b) = 1$ in case $b = 0$ and $o(\pi^b) = p$ in case $b = 1,2, \ldots, p-1$. Thus the only elements $\sigma_{1,b}$ in H satisfying $o(\sigma_{1,b}) = p$ are $\pi, \pi^2, \pi^3, \ldots, \pi^{p-1}$.

To complete the proof, we show that $a \neq 1$ implies $o(\sigma_{a,b}) \neq p$. If $a \neq 1$ and $\sigma = \sigma_{a,b}$, then

$$u\sigma^2 = a(au + b) + b = a^2u + (a + 1)b,$$

$$u\sigma^3 = (a^2u + (a + 1)b) + b = a^3u + (a^2 + a + 1)b$$

and similarly

$$u\sigma^n = a^nu + (a^{n-1} + a^{n-2} + \cdots + a^2 + a + 1)b$$

for any $n \in \mathbb{N}$. As $a - 1 \in \mathbb{F}_p^\times$, we can write

$$u\sigma^n = a^nu + \frac{a^n - 1}{a-1}b$$

from which we read that $\sigma^n = 1$ if and only if $a^n = 1$, so $o(\sigma) = o(a)$, the order of a in the multiplicative group $\mathbb{F}_p^\times$. But $\mathbb{F}_p^\times$ has order $p - 1$ and, by Lagrange's theorem, there is no a in $\mathbb{F}_p^\times$ with $o(a) = p$. Thus $\sigma_{a,b}$ cannot be of order p if $a \neq 1$.

(2) By hypothesis, $\langle\pi\rangle \leqslant H$. Let $\varphi \in Aut(H)$. Then $1 \neq \pi\varphi$ is an element of order p in H. Then $\pi\varphi = \pi^a$ for some $a \in \{1,2, \ldots, p-1\}$, so $\langle\pi\rangle\varphi = \langle\pi\varphi\rangle = \langle\pi^a\rangle = \langle\pi\rangle$ and $\langle\pi\rangle$ is characteristic and therefore also normal in H.

(3) By hypothesis, $\langle\pi\rangle \trianglelefteq H$. We must prove $H \leqslant A(p)$. Let $\sigma \in H$. Then $1 \neq \sigma^{-1}\pi\sigma = \pi^\sigma \in \langle\pi\rangle^\sigma = \langle\pi\rangle$ and $\pi^\sigma = \pi^a$ for some $a \in \{1,2, \ldots, p-1\}$. So $\pi\sigma = \sigma\pi^a$ and

$$(t+1)\sigma = t\pi\sigma = t\sigma\pi^a = t\sigma + a, \qquad \text{for any } t \in \{1,2,\ldots,p-1,p\}.$$

Then $(t+2)\sigma = (t+1)\sigma + a = t\sigma + 2a,$
$(t+3)\sigma = (t+2)\sigma + a = t\sigma + 3a,$

and similarly $(t+u)\sigma = t\sigma + ua$ for all $t,u \in \{1,2,\ldots,p-1,p\}$. Putting $t = 0$ and $t\sigma = b$, we get $u\sigma = t\sigma + ua = au + b$ for any $u = 1,2,\ldots,p-1,p$. Therefore $\sigma = \sigma_{a,b} \in A(p)$. This proves $H \leqslant A(p)$.

(4) Assume now H is a linear subgroup of S_p and $H \trianglelefteq K \leqslant S_p$, We must prove $K \leqslant A(p)$. Now $<\pi>$ is a characteristic subgroup of H by part (2), and $<\pi>$ is a normal subgroup of K by Lemma 23.15, so K is a linear subgroup of S_p by part (3). □

59.22 Lemma: *Let p be a prime number and K a transitive subgroup of S_p. If $1 \neq H \trianglelefteq K$, then H is also transitive.*

Proof: Let $i,j \in \{1,2,\ldots,p\}$. We claim that the number of elements in the H-orbit of i is equal to the number of elements in the H-orbit of j. Indeed, since K is transitive, there is a $\tau \in K$ with $i\tau = j$ and

$$|H\text{-orbit of } i| = |H:Stab_H(i)| = |H:Stab_K(i) \cap H| = |H^\tau:(Stab_K(i) \cap H)^\tau|$$

$$= |H^\tau:(Stab_K(i))^\tau \cap H^\tau| = |H:(Stab_K(i))^\tau \cap H| = |H:Stab_K(i\tau) \cap H|$$

$$= |H:Stab_K(j) \cap H| = |H:Stab_H(j)| = |H\text{-orbit of } j|$$

in view of Lemma 25.10 and Lemma 25.8. Thus all orbits of H have the same number of elements, say m. If k is the number of H-orbits, then the $\{1,2,\ldots,p\}$ is partitioned into k subsets each of which has m elements. Thus $p = mk$ and $k = p$ or $k = 1$. If $k = p$ were true, i.e., if there were p H-orbits, the H-orbits would consist of single terms and we would get $u\sigma = u$ for any $u \in \{1,2,\ldots,p\}$, $\sigma \in H$. This would give $H = 1$, contrary to the hypothesis. Hence $k = 1$ and H is transitive. □

59.23 Lemma: *Let p be a prime number and $G \leqslant S_p$. Then G is transitive if and only if p divides the order of G.*

Proof: If $p \mid |G|$, there is an element σ of G with $o(\sigma) = p$. Then σ is a cycle of length p, say $(a_1 a_2 \ldots a_p)$. Then any a_i is mapped to any a_j by $\sigma^{|j-i-1|} \in G$ and so G is transitive. Conversely, if G is transitive, there is, for each $a = 1,2,\ldots,p$, a permutation σ_a with $1\sigma_a = a$ and we have the coset decomposition

$$G = \bigcup_{i=1}^{p} [Stab_G(1)]\sigma_a,$$

whence $|G| = |Stab_G(1)|p$ is divisible by p. □

We can now find all solvable transitive subgroups of S_p. Basically, we use Lemma 59.21 and Lemma 59.22 to go downwards and upwards along a composition series of such subgroups.

59.24 Theorem: *Let p be a prime number and $G \leqslant S_p$. Then G is a solvable transitive subgroup of S_p if and only if G is conjugate to a linear subgroup of S_p.*

Proof: Let G be a solvable transitive subgroup of S_p. Consider a composition series of G, say

$$1 = H_0 \lhd H_1 \lhd H_2 \lhd \ldots \lhd H_{m-1} \lhd H_m = G.$$

The composition factors H_i/H_{i-1} are cyclic of prime order by Theorem 27.18. Since H_m is transitive, H_{m-1} is also transitive by Lemma 59.22, and then H_{m-2} is transitive, then H_{m-3} transitive and so on. In this way, we see that H_1 is transitive. Then p divides $|H_1|$ by Lemma 59.23 and we get $|H_1| = p$. So H_1 is a cyclic group generated by a cycle $(a_1 a_2 \ldots a_p)$. Replacing G by a conjugate of G, we may assume $H_1 = \langle\pi\rangle = \langle(12\ldots p)\rangle$. Now Lemma 59.21(4) shows that H_2 is a linear subgroup of S_p, so H_3 is also a linear subgroup of S_p, so H_4 is also a linear subgroup of S_p and so on. In this way, we conclude $H_m = G$ is a linear subgroup of S_p.

Conversely, let G be a linear subgroup of S_p. Then π is a subgroup of G and so p divides $|G|$ and Lemma 59.23 shows G is a transitive subgroup of S_p. Now we have to prove G is solvable. As $G \leqslant A(p)$, it will be sufficient to prove that $A(p)$ is solvable. In view of the multiplication rule $\sigma_{a,b}\sigma_{c,d} = \sigma_{ac,bc+d}$, the mapping

$$\varphi: A(p) \to \mathbb{F}_p^*$$
$$\sigma_{a,b} \to a$$

is a homomorphism onto $\mathbb{F}_p^*$, with $Ker\ \varphi = \{\sigma_{1,b} \in A(p): b \in \mathbb{F}_p\} = <\pi>$. Thus $<\pi> \trianglelefteq A(p)$ and $A(p)/<\pi> = A(p)/Ker\ \varphi \cong Im\ \varphi = \mathbb{F}_p^*$ is abelian. Then

$$1 \trianglelefteq <\pi> \trianglelefteq A(p)$$

is an abelian series of $A(p)$ and hence $A(p)$ is solvable. □

We give another group theoretical characterization of solvable transitive subgroups of S_p. This will be translated into Galois' characterization of polynomial equations of prime degree which are solvable by radicals.

59.25 Theorem: *Let p be a prime number and G a transitive subgroup of S_p. Then G is solvable if and only if ι is the only permutation in G that fixes two numbers from $\{1,2, \ldots,p\}$, i.e., if and only if*

$$Stab_G(i) \cap Stab_G(j) = 1$$

for any two distinct i,j from $\{1,2, \ldots,p\}$.

Proof: Suppose first that G is a solvable transitive subgroup of S_p. Then G is conjugate to a subgroup of $A(p)$, say $G = H^\tau$, where $<\pi> \leqslant H \leqslant A(p)$ and $\tau \in S_p$ (Theorem 59.24). If i,j are distinct numbers from $\{1,2, \ldots,p\}$ and $\sigma \in G$ fixes both i and j, then $\sigma^{\tau^{-1}} \in G^{\tau^{-1}} = (H^\tau)^{\tau^{-1}} = H \leqslant A(p)$ fixes both $i\tau^{-1}$ and $j\tau^{-1}$. But, aside from the identity, there is no permutation in $A(p)$ that fixes two distinct numbers from $\{1,2, \ldots,p\}$. Thus $\sigma^{\tau^{-1}} = \iota$ and $\sigma = \iota$. So the identity permutation is the only permutation in G that fixes two numbers in $\{1,2, \ldots,p\}$.

Now suppose conversely that G is a transitive subgroup of S_p with the property that the identity is the only permutation in G that fixes two numbers in $\{1,2, \ldots,p\}$. Let i,j be two distinct numbers in $\{1,2, \ldots,p\}$ and write

$$H = Stab_{S_p}(i) \cap Stab_{S_p}(j) = \{\tau \in S_p : i\tau = i \text{ and } j\tau = j\}.$$

The hypothesis gives $H \cap G = 1$. If $\sigma_1,\sigma_2 \in G$ and σ_1,σ_2 belong to the same right coset of H in S_p, then $\sigma_1\sigma_2^{-1}$ belongs to $H \cap G = 1$, so $\sigma_1 = \sigma_2$. Thus

there is at most one element of G in each right coset of H in S_p. So $|G|$ is less than or equal to the number $|S_p:H|$ of right cosets of H in S_p and, as H is isomorphic to S_{p-2}, we have $|G| \leq |S_p:H| = |S_p|/|H| = p!/(p-2)! = p(p-1)$. Lemma 59.23 yields p divides $|G|$, so there is an element $\pi' = (a_1 a_2 \ldots a_p)$ of order p in G. If we write $\tau = \begin{pmatrix} a_1 a_2 \ldots a_p \\ 1\,2 \ldots p \end{pmatrix} \in S_p$, then $(12\ldots p) = \pi = \pi'^{\tau}$ is an element of order p in G^{τ}. Aside from the powers of π, there is no permutation of order in G^{τ}, for if $\sigma \in G^{\tau}$ had order p and $\langle\pi\rangle \cap \langle\sigma\rangle \neq 1$, then $|\langle\pi\rangle \cap \langle\sigma\rangle| = |\langle\pi\rangle||\langle\sigma\rangle| = p^2$ (Lemma 19.6) and so there would be at least p^2 distinct elements in G^{τ}, whereas $|G^{\tau}| = |G|$ is at most $p^2 - p$. So $\langle\pi\rangle$ is a normal subgroup of G^{τ} by Lemma 59.21(2) and G^{τ} is a linear subgroup of S_p by Lemma 59.21(3). Hence G is conjugate to a linear subgroup of S_p and G is solvable by Theorem 59.24. □

For the sake of completeness, we prove Galois' theorem stating that a polynomial equation of prime degree is solvable by radicals if and only if "all roots can be expressed rationally in terms of any two of them."

59.26 Theorem: *Let K be a field of characteristic 0 and let $f(x)$ be an irreducible polynomial of prime degree p in $K[x]$. The equation $f(x) = 0$ is solvable by radicals if and only if, for any two distinct roots a,b of $f(x)$, $K(a,b)$ is a splitting field of $f(x)$ over K.*

Proof: Let E be a splitting field of $f(x)$ over K and G the Galois group of $f(x)$. Then E is a Galois extension of K (Theorem 55.7) and G is a transitive subgroup of S_p (Theorem 56.17). Let a,b be two distinct roots of $f(x)$ and let $J = K(a,b)'$ be the subgroup of G corresponding to it.

$$\begin{array}{rcl} E & & 1 \\ K(a,b) & & J \\ K & & G \end{array}$$

J is the subgroup of G consisting precisely of the permutations of the roots fixing a and b. Now we have the is equivalences

$$E = K(a,b) \iff J = 1$$

$\Leftrightarrow$ 1 is the only permutation in G fixing a and b

$\Leftrightarrow$ G is a solvable subgroup of S_p

$\Leftrightarrow$ the equation $f(x) = 0$ is solvable by radicals.

This completes the proof. □

*

* *

In this part, we give algebraic formulas for the roots of polynomials of degree two, three and four. For the sake of generality, we assume the coefficients are indeterminates, but, as will be clear from the arguments, the formulas are valid if the coefficients are taken from the base field.

59.27 Theorem: *Let K be a field with char $K \neq 2$ and let a,b be indeterminates over K so that*

$$g(x) = x^2 - ax + b$$

is the general polynomial of degree two over K. Then the roots r_1, r_2 of $g(x)$ are given by

$$r_1 = \frac{a+\sqrt{D}}{2}, \qquad r_2 = \frac{a-\sqrt{D}}{2},$$

where $D = a^2 - 4b$.

Proof: The discriminant D of $g(x)$ is $(r_1 - r_2)^2 = (r_1 + r_2)^2 - 4r_1r_2 = a^2 - 4b$. Hence $r_1 + r_2 = a$ and $r_1 - r_2 = \sqrt{D}$. Solving this system of linear equations for r_1, r_2, we find

$$r_1 = \frac{a+\sqrt{D}}{2}, \qquad r_2 = \frac{a-\sqrt{D}}{2}. \qquad \square$$

In particular, the cubic roots of unity, which are the roots of the polynomial $x^2 + x + 1$, are given by $\alpha_1 = (-1 + \sqrt{-3})/2$, $\alpha_2 = (-1 - \sqrt{-3})/2$. This will be used in the next theorem.

59.28 Theorem: *Let K be a field with $\operatorname{char} K \neq 2,3$ and assume that K contains a primitive cube root of unity, say α. Let a,b,c be distinct indeterminates over K and let*

$$g(x) = x^3 - ax^2 + bx - c$$

be the general cubic polynomial over K. Then the roots r_1,r_2,r_3 of $g(x)$ are given by

$$r_1 = \frac{1}{3}(a + u + v) \qquad r_2 = \frac{1}{3}(a + \alpha^2 u + \alpha v) \qquad r_2 = \frac{1}{3}(a + \alpha u + \alpha^2 v),$$

where

$$u = \sqrt[3]{a^3 - \frac{9}{2}ab + \frac{27}{2}c + \frac{3}{2}\sqrt{-3D}}$$

$$v = \sqrt[3]{a^3 - \frac{9}{2}ab + \frac{27}{2}c - \frac{3}{2}\sqrt{-3D}}$$

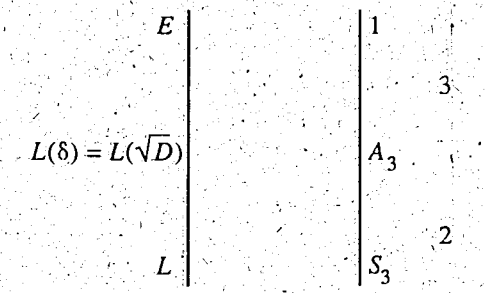
are such that $uv = a^2 - 3b$ and $D = \frac{4}{27}(a^2 - 3b)^3 - \frac{1}{27}(2a^3 - 9ab + 27c)^2$.

Proof: Let $E = K(r_1,r_2,r_3)$ be a splitting field of $g(x)$ over $L = K(a,b,c)$. Then E is a Galois extension of L and the Galois group $Aut_L E$ of $g(x)$ is S_3 (Theorem 59.17). Since S_3 is transitive, $g(x)$ is irreducible over L, and $\operatorname{char} K \neq 3$ implies that the derivative of $g(x)$ is not zero, so $g(x)$ has no common root with its derivative and the roots r_1,r_2,r_3 are distinct. Under the Galois correspondence, the alternating group A_3 corresponds the subfield $L(\delta)$ of E, where $\delta = (r_1 - r_2)(r_1 - r_3)(r_2 - r_3)$ is the square root of the discriminant of $g(x)$ (Theorem 56.18, Theorem 56.19).

Let D be the discriminant of $g(x)$. We evaluate D. Observe that $r_1 - (a/3)$, $r_2 - (a/3), r_3 - (a/3)$ are the roots of $g(x + (a/3))$, so the root differences and the discriminant of $g(x)$ are the same as those of $g(x + (a/3))$. One obtains easily $g(x + (a/3)) = x^3 + px + q$, where $p = (3b - a^2)/3$ and $q = (-2a^3 + 9ab - 27c)/27$. Then $D = -4p^3 - 27q^2$ is computed to be

$$\frac{4}{27}(a^2 - 3b)^3 - \frac{1}{27}(2a^3 - 9ab + 27c)^2$$

(Example 56.10(b)).



Now E is a cyclic extension of $L(\sqrt{D})$, because E is Galois over $L(\sqrt{D})$ (Theorem 54.25(1)) and its Galois group A_3 is cyclic of order 3. Thus E is obtained by adjoining a root u of a polynomial $x^3 - h \in L(\sqrt{D})[x]$ to $L(\sqrt{D})$ (Theorem 57.11). A generator σ of A_3 maps as follows

$$\begin{array}{lll} r_1 \to r_2, & r_2 \to r_3, & r_3 \to r_1; \\ u \to \alpha u, & \alpha u \to \alpha^2 u, & \alpha^2 u \to u. \end{array}$$

An examination of the proof of Theorem 57.11 reveals that u should be taken as a nonzero element of the form $\alpha d + \alpha(\alpha\sigma)d\sigma + \alpha(\alpha\sigma)(\alpha\sigma^2)d\sigma^2$, with $d \in E$. So we must find a $d \in E$ such that $d + \alpha d\sigma + \alpha^2 d\sigma^2 \neq 0$. We choose $d = r_1$. So let $u = r_1 + \alpha r_2 + \alpha^2 r_3$. Similarly we put $v = r_1 + \alpha^2 r_2 + \alpha r_3$.

We already know $u^3 \in L(\sqrt{D})$. We now evaluate it. We have

$$u^3 = (r_1 + \alpha r_2 + \alpha^2 r_3)^3 = r_1^3 + r_2^3 + r_3^3 + 3\alpha A + 3\alpha^2 B + 6r_1 r_2 r_3, \qquad (1)$$

where we put $A = r_1^2 r_2 + r_2^2 r_3 + r_3^2 r_1$ and $B = r_1 r_2^2 + r_2 r_3^2 + r_3 r_1^2$ for shortness. The method of §38 gives

$$r_1^3 + r_2^3 + r_3^3 = (r_1 + r_2 + r_3)^3 - 3(A + B) - 6r_1 r_2 r_3,$$

and
$$\begin{aligned} A + B &= r_1^2 r_2 + r_2^2 r_3 + r_3^2 r_1 + r_1 r_2^2 + r_2 r_3^2 + r_3 r_1^2 \\ &= (r_1 + r_2 + r_3)(r_1 r_2 + r_1 r_3 + r_2 r_3) - 3r_1 r_2 r_3 = ab - 3c, \end{aligned}$$

$$\begin{aligned} A - B &= r_1^2 r_2 + r_2^2 r_3 + r_3^2 r_1 - r_1 r_2^2 - r_2 r_3^2 - r_3 r_1^2 \\ &= (r_1 - r_2)(r_1 - r_3)(r_2 - r_3) = \sqrt{D}, \end{aligned}$$

so (1) becomes

$$u^3 = [(r_1 + r_2 + r_3)^3 - 3[(\frac{ab-3c+\sqrt{D}}{2}) + (\frac{ab-3c-\sqrt{D}}{2})] - 6r_1r_2r_3]$$

$$+3(\frac{-1+\sqrt{-3}}{2})(\frac{ab-3c+\sqrt{D}}{2}) + 3(\frac{-1-\sqrt{-3}}{2})(\frac{ab-3c-\sqrt{D}}{2}) + 6r_1r_2r_3$$

$$= a^3 - \frac{9}{2}ab + \frac{27}{2}c + \frac{3}{2}\sqrt{-3D}$$

and a similar calculation yields

$$v^3 = a^3 - \frac{9}{2}ab + \frac{27}{2}c - \frac{3}{2}\sqrt{-3D}.$$

So u and v are cube roots of the expressions found above. But there are three cube roots of these expressions, and we must decide which cube roots we should take. This is found from

$$uv = (r_1 + \alpha r_2 + \alpha^2 r_3)(r_1 + \alpha^2 r_2 + \alpha r_3) = a_1^2 + a_2^2 + a_3^2 - a_1a_2 - a_1a_3 - a_2a_3$$
$$= a^2 - 3b.$$

The cube roots must be therefore so chosen that their product will be equal to $a^2 - 3b$. If

$$u = \sqrt[3]{a^3 - \frac{9}{2}ab + \frac{27}{2}c + \frac{3}{2}\sqrt{-3D}}$$

$$v = \sqrt[3]{a^3 - \frac{9}{2}ab + \frac{27}{2}c - \frac{3}{2}\sqrt{-3D}}$$

are denote cube roots with this property, then, solving the equations

$$a = r_1 + r_2 + r_3$$
$$u = r_1 + \alpha r_2 + \alpha^2 r_3$$
$$v = r_1 + \alpha^2 r_2 + \alpha r_3$$

for r_1, r_2, r_3, we get

$$r_1 = \frac{1}{3}(a + u + v) \qquad r_2 = \frac{1}{3}(a + \alpha^2 u + \alpha v) \qquad r_2 = \frac{1}{3}(a + \alpha u + \alpha^2 v),$$

as was to be proved. □

A remarkable fact is that, if $f(x) \in \mathbb{R}[x]$ has three real roots, then the roots of $f(x)$ cannot be expessed in terms of real radicals. We want to dicsuss this matter. We need an elementary lemma.

59.29 Lemma: *Let K be a field, $a \in K$ and p a prime number. Assume $\mathrm{char}\, K \neq p$. If $x^p - a \in K[x]$ is reducible in $K[x]$, then $a = c^p$ for some $c \in K$.*

Proof: In a splitting field of $x^p - a$ over K, we have the decomposition

$$x^p - a = \prod_{k=0}^{p-1} (x - \zeta^k u)$$

where u is a root of $x^p - a$ and ζ is a primitive p-th root of unity. If $x^p - a$ is reducible in $K[x]$ and $f(x) \in K[x]$ is a factor of $x^p - a$ with $1 < \deg f(x) < p$, then $f(x)$ is a product of some of the $x - \zeta^k u$, and the constant term $(-1)^h b_0$ of $f(x)$ is $(-1)^h \zeta^m u^h$ for some $m \in \mathbb{N}$, where $h = \deg f(x)$. So $b_0 = \alpha u^h$ for some p-th root of unity, so ${b_0}^p = u^{ph} = a^h$ and, since $(h,p) = 1$, there are integers k,n satisfying $kh + np = 1$. Thus $a = a^{kh}a^{np} = ({b_0}^p)^k a^{hp} = ({b_0}^k a^n)^p$ and ${b_0}^k a^n \in K$. □

59.30 Lemma: *Let K be a subfield of $\mathbb{R}$ and $f(x)$ an irreducible cubic polynomial in $K[x]$. Let S be a splitting field of $f(x)$ over K. If $f(x)$ has three distinct real roots, then there is no radical extension R of K such that $S \subseteq R \subseteq \mathbb{R}$.*

Proof: Let r_1,r_2,r_3 be the roots and $D = (r_1 - r_2)^2(r_1 - r_3)^2(r_2 - r_3)^2$ the discriminant of $f(x)$. Then D is a positive real number. Put $K_1 = K(\sqrt{D}) \subseteq \mathbb{R}$. Clearly K_1 is a subfield of S. We may assume that $f(x)$ is monic.

Suppose, by way of contradiction, there is a radical extension R of K with $S \subseteq R \subseteq \mathbb{R}$. Then RK_1 is a radical extension of K_1 (Lemma 59.4). So there is a finite chain of fields

$$K_1 \subseteq K_2 \subseteq \ldots \subseteq K_{n-1} \subseteq K_n = RK_1$$

such that $K_i = K_{i-1}(u_i)$ for some root u_i of a polynomial of the form $x^{m_i} - a_i$ in $K_{i-1}[x]$ $(i = 2,3,\ldots,n)$. We may assume m_i are prime numbers. Moreover, after deleting redundant fields, we may assume $u_i \notin K_{i-1}$. Thus we

assume m_i are prime, $u_i^{m_i} \in K_{i-1}$ and $u_i \notin K_{i-1}$. Then $x^{m_i} - u_i^{m_i} \in K_{i-1}[x]$ is irreducible in $K_{i-1}[x]$, for otherwise we had $u_i^{m_i} = c^{m_i}$ for some $c \in K_{i-1}$ (Lemma 59.29) and u_i/c, which is distinct from ∓ 1 in view of $u_i \notin K_{i-1}$, would be a primitive m_i-th root of unity, so $u_i/c \in K_{i-1} \subseteq \mathbb{R}$ would be complex number with nonzero imaginary part, a contradiction. Therefore $x^{m_i} - u_i^{m_i} \in K_{i-1}[x]$ is irreducible over K_{i-1} and is in fact the minimal polynomial of $u_i \in K_i$ over K_{i-1}. This gives $|K_i:K_{i-1}| = m_i$.

$f(x)$ is irreducible in $K_1[x]$, for $f(x)$ is the minimal polynomial of any of its roots over K and if r_1, say, were in K_1, then $2 = |K_1:K| = |K_1:K(r_1)||K(r_1):K|$ would be divisible by $|K(r_1):K| = deg\ f(x) = 3$, which is nonsense. Now S is a splitting field of $f(x)$ over K_1 (Example 53.5(e)) and since $\sqrt{D} \in K_1$, the Galois group $Aut_{K_1} S$ is isomorphic to A_3. (Theorem 56.21).

On the other hand, the roots of $f(x)$ are in $S \subseteq R \subseteq R K_1$ and $f(x)$ is reducible over $R K_1 = K_n$. Let K_i be the field in the chain above where $f(x)$ becomes reducible, that is to say, let $i \in \{2, \ldots, n\}$ be such that $f(x)$ is irreducible over K_{i-1} and reducible over $K_i = K_{i-1}(u_i)$. Then there is a root of $f(x)$ in K_i, say $r_1 \in K_i$ and, as above, $f(x)$ is the minimal polynomial of r_1 over K_{i-1}, so the prime number $m_i = |K_i:K_{i-1}| = |K_i:K_{i-1}(r_1)|\,|K_{i-1}(r_1):K_{i-1}|$ is divisible by $|K_{i-1}(r_1):K_{i-1}| = deg\ f(x) = 3$ and so $m_i = 3$. Thus K_i is an extension of K_{i-1} containing the root r_1 of $f(x)$.

Let N be a splitting field of $f(x)$ over K_{i-1}. Then N/K_{i-1} is a Galois extension (Theorem 55.7) and since $\sqrt{D} \in K_{i-1}$, the Galois group $Aut_{K_{i-1}} N$ is isomorphic to A_3 (Theorem 56.21). So $|N:K_{i-1}| = |Aut_{K_{i-1}} N| = 3$. From $r_1 \notin K_{i-1}$ and $r_1 \in N \cap K_i$, we get $K_{i-1} \subset N \cap K_i \subseteq N$ and degree considerations force $N \cap K_i = N$, so $N \subseteq K_i$ and as $|N:K_{i-1}| = 3 = |K_i:K_{i-1}|$, we deduce $N = K_i$.

Theorem 55.10 yields now that K_i is normal over K_{i-1} and since the irreducible polynomial $x^{m_i} - a_i$ in $K_{i-1}[x]$ has a root u_i in K_i, the other roots $u_i\omega$ and $u_i\omega^2$ of $x^{m_i} - a_i$ are in K_i, so $\omega = u_i\omega/u_i \in K_i$. This contradicts $K_i \subseteq \mathbb{R}$.

Thus there can be no radical extension R of K such that $S \subseteq R \subseteq \mathbb{R}$. □

59.31 Theorem: *Let K be a field with $\operatorname{char} K \neq 2,3$ and assume that K contains a primitive cube root of unity. Let a,b,c,d be distinct indeterminates over K and let*

$$g(x) = x^4 - ax^3 - bx^2 + cx - d$$

be the general biquadratic polynomial over K. Then the roots r_1,r_2,r_3,r_4 of $g(x)$ are given by

$$r_1 = \frac{1}{4}(a + \sqrt{u} + \sqrt{v} + \sqrt{y})$$

$$r_2 = \frac{1}{4}(a + \sqrt{u} - \sqrt{v} - \sqrt{y})$$

$$r_3 = \frac{1}{4}(a - \sqrt{u} + \sqrt{v} - \sqrt{y})$$

$$r_4 = \frac{1}{4}(a - \sqrt{u} - \sqrt{v} + \sqrt{y}),$$

where $u = a^2 - 4\beta - 4\gamma$, $v = a^2 - 4\alpha - 4\gamma$, $y = a^2 - 4\alpha - 4\beta$,

α,β,γ are the roots of $x^3 - bx^2 + (ac - 4d)x - (a^2d - 4bd + c^2)$ and the square roots are subject to the condition

$$\sqrt{u}\sqrt{v}\sqrt{y} = -a^3 + 4ab - 8c.$$

Proof: Let r_1,r_2,r_3,r_4 be the roots of $g(x)$ and let $\alpha = r_1r_2 + r_3r_4$, $\beta = r_1r_3 + r_2r_4$, $\gamma = r_1r_4 + r_2r_3$. Then α,β,γ are the roots of the resolvent cubic

$$x^3 - bx^2 + (ac - 4d)x - (a^2d - 4bd + c^2)$$

and the Galois group of $g(x)$, regarded as a polynomial in $K(\alpha,\beta,\gamma)$ is V_4 (Theorem 56.23, Theorem 59.17, Lemma 56.25). We can solve for α,β,γ in terms of radicals by the method of Theorem 59.28. As $|V_4| = 4$, we can find the roots r_1,r_2,r_3,r_4 by introducing two square roots. For this purpose, we put

$$u = (r_1 + r_2 - r_3 - r_4)^2$$
$$v = (r_1 - r_2 + r_3 - r_4)^2$$
$$y = (r_1 - r_2 - r_3 + r_4)^2.$$

An easy computation gives

$$u = a^2 - 4\beta - 4\gamma;\ v = a^2 - 4\alpha - 4\gamma;\ y = a^2 - 4\alpha - 4\beta$$

and $(r_1 + r_2 - r_3 - r_4)(r_1 - r_2 + r_3 - r_4)(r_1 - r_2 - r_3 + r_4)$ is a symmetric polynomial in the roots r_1, r_2, r_3, r_4, found easily to be $-a^3 + 4ab - 8c$. Hence we have

$$\begin{aligned} a &= r_1 + r_2 + r_3 + r_4 \\ \sqrt{u} &= r_1 + r_2 - r_3 - r_4 \\ \sqrt{v} &= r_1 - r_2 + r_3 - r_4 \\ \sqrt{y} &= r_1 - r_2 - r_3 + r_4, \end{aligned}$$

provided we choose the square roots in such a way that $\sqrt{u}\sqrt{v}\sqrt{y} = -a^3 + 4ab - 8c$. Solving this system of linear equations, we find

$$r_1 = \frac{1}{4}(a + \sqrt{u} + \sqrt{v} + \sqrt{y}), \qquad r_2 = \frac{1}{4}(a + \sqrt{u} - \sqrt{v} - \sqrt{y}),$$

$$r_3 = \frac{1}{4}(a - \sqrt{u} + \sqrt{v} - \sqrt{y}), \qquad r_4 = \frac{1}{4}(a - \sqrt{u} - \sqrt{v} + \sqrt{y}). \qquad \square$$

*

* *

In this part, we settle some famous problems.

A real number a will be called *constructible* if it is possible to draw a line segment of length $|a|$ using ruler and compass only in a finite number of steps. Thus "constructible" means "constructible by ruler and compass". Similarly, "to draw" will mean "to draw using ruler and compass only". Each step in a ruler and compass construction is one of the following types:

(i) finding the intersection point of two straight lines;

(ii) finding the intersection points of two circles;

(iii) finding the intersection points of a straight line and a circle.

From elementary geometry, it is known that, for any given line l and a given P, we can draw a line through P parallel to l and also a line through P perpendicular to l.

We draw two perpendicular lines and regard them as coordinate axes. Then we fix a unit length. Then we can draw line segments on the coordinate axes with integral length. Since we can draw lines parallel and/or perpendicular to the axes, we can locate all points in the Euclidean plane with integer coordinates as intersection of lines paralel to the axes.

After the introduction of a coordinate system on the plane, we see that a real number $a \neq 0$ is constructible if and only if the line segment $[0,a] \times \{0\}$ or $\{0\} \times [0,a]$ is constructible. (Closed intervals. Here $[0,a]$ is to be read as $[a,0]$ when $a < 0$.)

Assume a and b are constructible. Then $a + b$ and $a - b$ are constructible, too. In addition, we can draw the line through $(a,0)$ parallel to the line segment joining $(0,b)$ and $(1,0)$, which intersects the y-axis at $(0,ab)$. Also, if $b \neq 0$, we can draw the line through $(0,1)$ parallel to the line segment joining $(0,b)$ and $(a,0)$, which intersects the y-axis at $(0,a/b)$. Thus $a + b$, $a - b$, ab and a/b are constructible whenever a and b are constructible ($b \neq 0$ in case of division). Thus the constructible real numbers form a subfield of $\mathbb{R}$. In particular, all rational numbers are constructible, for $\mathbb{Q}$ is the prime subfield of the field of constructible real numbers.

A point (a,b) is said to be *constructible* if both a and b are constructible. This is the case if and only if (a,b) can be determined by a finite sequence of ruler and compass constructions starting from points with integer coordinates. Hence all points with rational coordinates are constructible.

In order to determine which numbers are constructible, i.e., in order to determine the coordinates of constructible points, we must examine what type of points arise after each of the ruler and compass construction steps (i),(ii),(iii). It will be convenient to introduce some terminology.

If K is a subfield of real numbers, a point (a,b) in the plane is called a *K-point* if both a and b are elements of K. A straight line through two distinct K-points is called a *K-line*. A circle whose center is an K-point and whose radius is an element of K is called a *K-circle*.

A K-line l has an equation of the form $ax + by + c = 0$, where $a,b,c \in K$, for if l is the straight line through the K-points (x_0,y_0) and (x_1,y_1), then l has the equation $(y_1 - y_0)x + (x_0 - x_1)y + (x_1y_0 - y_1x_0) = 0$. A K-circle C has an equation of the form $x^2 + y^2 + ax + by + c = 0$, where $a,b,c \in K$, for if the center of C is the K-point (x_0,y_0) and the radius of C is the K-number r, then C has the equation $x^2 + y^2 + (-2x_0)x + (-2y_0)y + (x_0^2 + y_0^2 - r^2) = 0$.

Now let K be a subfield of $\mathbb{R}$. We determine the nature of intersection points of two K-two straight lines and/or K-circles that arise as a result of one of the steps (i),(ii),(iii).

If l and m are K-lines, with equations $ax + by + c = 0$ and $dx + ey + f = 0$, say, where $a,b,c,d,e,f \in K$, then l and m intersect if and only if $ae - bd \neq 0$ and their point of intersection can be found, on solving the system of linear equations

$$ax + by = -c$$
$$dx + ey = -f$$

for x,y by Cramer's rule, to be the K-point

$$((-ce + bf)/(ae - bd),\ (-af + cd)/(ae - bd))$$

(Theorem 45.2). Thus two K-lines intersect (if at all) at a K-point.

Let C be a K-circle and l a K-line, with equations $x^2 + y^2 + ax + by + c = 0$ and $dx + ey + f = 0$, say, where $a,b,c,d,e,f \in K$. We find the intersection points (x,y) of of C and l. Here d,e cannot both be 0, for then the equation $dx + ey + f = 0$ would not represent a straight line. In case $d = 0$, we have $e \neq 0$ and $y = -f/e$, so $x^2 + (-f/e)^2 + ax + b(-f/e) + c = 0$, giving $Ax^2 + Bx + E = 0$, with $A,B,E \in K$. Hence the x-coordinate of an intersection point of C and l is a root of a quadratic polynomial over K. Let D be the discriminant of this polynomial. Either $D < 0$ and this polynomial has no real roots, so C and l do not intersect; or $D \geqslant 0$ and it has two (possibly equal) roots x_1,x_2 in the field $K(\sqrt{D})$, so C and l intersect at two $K(\sqrt{D})$-points $(x_1,-f/e)$, $(x_2,-f/e)$. In case $d \neq 0$, we put $g = -e/d$, $h = -f/d$ and use the equation $x = gy + h$ of l. Here $g,h \in K$. Now $(gy + h)^2 + y^2 + a(gy + h) + by + c = 0$ gives $Ay^2 + By + E = 0$, with $A,B,E \in K$ (not the same A,B,E as above). Hence the y-coordinate of an intersection point of C and l is a root of a quadratic polynomial over K. Let D be the discriminant of this polynomial. Either $D < 0$ and this polynomial has no real roots, so C and l do not intersect; or $D \geqslant 0$ and it has two (possibly equal) real roots y_1,y_2 in the field $K(\sqrt{D})$,

so C and l intersect at two (possibly identical) $K(\sqrt{D})$-points $(gy_1 + h, y_1)$, $(gy_2 + h, y_2)$.

Let C_1, C_2 be K-circles, say with equations $x^2 + y^2 + ax + by + c = 0$ and $x^2 + y^2 + dx + ey + f = 0$, where $a,b,c,d,e,f \in K$. Then the intersection points of C_1 and C_2 are the same as the intersection points of C_1 and the K-line $(a - d)x + (b - e)y + c = 0$. Thus either C_1, C_2 do not intersect or they intersect at two (possibly identical) $K(\sqrt{D})$-points, where $D \in K$.

So each step in a ruler and compass construction gives rise to a K-point or a $K(\sqrt{D})$-point for some $D \in K$, if K denotes the field of lines/circles used in that step.

A real number a is constructible if and only if the point $(a,0)$ is constructible, hence if and only if the point $(a,0)$ can be obtained as a result of a finite sequence of the steps (i),(ii),(iii) beginning with points having rational coordinates. Thus a is constructible if and only if there is a finite chain of fields

$$\mathbb{Q} = K_0 \subseteq K_1 \subseteq K_2 \subseteq \ldots \subseteq K_{n-1} \subseteq K_n$$

such that $K_i = K_{i-1}(\sqrt{D_i})$ for some $D_i \in K_i$ and $a \in K_n$. Here $|K_i:K_{i-1}| = 1$ or 2 according as $\sqrt{D_i}$ is or is not in K_{i-1}; hence $|K_n:\mathbb{Q}|$ is a power of two. Moreover $a \in K_n$, so $\mathbb{Q} \subseteq \mathbb{Q}(a) \subseteq K_n$ and $|\mathbb{Q}(a):\mathbb{Q}|$, being a divisor of $|K_n:\mathbb{Q}|$, is also a power of two. Thus a is algebraic over $\mathbb{Q}$ and the degree of a is a power of two. We proved the following theorem.

59.32 Theorem: If a real number a is constructible, then a is algebraic over $\mathbb{Q}$ and the degree of a over $\mathbb{Q}$ is a power of two. □

The converse of Theorem 59.32 is also true. See Ex. 10. We are now in a position to resolve some famous construction problems. The first one is the construction of a cube whose volume is twice the volume of a given cube (duplication of a cube). Choosing the length of a side of the given cube as unit length, the side of the cube to be constructed has length $\sqrt[3]{2}$. Thus the problem is to construct the real number $\sqrt[3]{2}$. Its minimal

polynomial is $x^3 - 2$, since this polynomial is irreducible over $\mathbb{Q}$ by Eisenstein's criterion. Thus $\sqrt[3]{2}$ is algebraic over $\mathbb{Q}$, but its degree over $\mathbb{Q}$ is three, not a power of two. Hence $\sqrt[3]{2}$ cannot be constructed: *it is impossible to duplicate a cube by ruler and compass alone.*

The second problem is to divide a given angle into three equal parts (trisection of an angle). An angle of θ radians is the circular arc of lenth θ on the unit circle, which we may assume to issue from the point (1,0) and terminate at the point $(\cos\theta, \sin\theta)$. It is constructible if and only if $(\cos\theta, \sin\theta)$ is constructible. In view of $\sin\theta = \mp\sqrt{1 - (\cos\theta)^2}$, we see that an angle of θ radians is constructible if and only if $\cos\theta$ is constructible. The problem is thus equivalent to: given $\cos\varphi$, construct $\cos(\varphi/3)$. From the trigonometric identity

$$\cos 3\theta = 4\cos^3\theta - 3\cos\theta,$$

we get, on writing a for $\cos\varphi$, the polynomial equation

$$4x^3 - 3x - a = 0$$

for $\cos(\varphi/3)$. The polynomial $4x^3 - 3x - a \in \mathbb{Q}(a)[x]$, where a is an indeterminate over $\mathbb{Q}$, is known as the *angle trisection polynomial.* It is irreducible over $\mathbb{Q}(a)$: to prove this, it will be sufficient to prove that it is irreducible over $\mathbb{Z}[a]$ (Lemma 34.11); but $4x^3 - 3x - a \in \mathbb{Z}[a][x] = \mathbb{Z}[x][a]$ is certainly irreducible in $\mathbb{Z}[x][a]$, because it is of degree one in a and its coefficients $(4x^3 - 3x), -1 \in \mathbb{Z}[x]$ are relatively prime in $\mathbb{Z}[x]$. So the angle trisection polynomial is irreducible over $\mathbb{Q}(a)$. The general trisection problem is whether it is possible to construct a root of

$$4x^3 - 3x - a = 0, \qquad a = \cos\varphi$$

in such a way that the construction remains valid when a is treated as an indeterminate. Since a root of the angle trisection polynomial has degree three over $\mathbb{Q}(a)$, the answer is negative: *it is impossible to find a method for trisecting an arbitrary angle by ruler and compass alone.* This does not mean of course that no *specific* angle can be trisected. On the contrary, there are angles like 90° that can very well be trisected.

The third problem is to draw a square whose area is the area of a given circle (squaring the circle). Choosing the radius of the given circle as unit

length, the side of the square to be constructed has length $\sqrt{\pi}$. Thus the problem is to construct the real number $\sqrt{\pi}$. But π and all the more so $\sqrt{\pi}$ are not algebraic over $\mathbb{Q}$ (Example 49.8(d)), let alone be of degree a power of two. Hence $\sqrt{\pi}$ cannot be constructed: *it is impossible to square the circle by ruler and compass alone.*

The final problem is to draw a regular n-gon. This is the same problem as dividing the circle into n equal parts. Thus we are to divide the angle of 2π radians into n equal parts, which means we are to construct the number $\cos(2\pi/n)$. Now $\cos(2\pi/n) = \dfrac{\zeta + \zeta^{-1}}{2} \in \mathbb{R}$, where $\zeta = e^{2\pi i/n} \in \mathbb{C}$ is a primitive n-th root of unity. The field $\mathbb{Q}(\cos(2\pi/n))$ is fixed only by the automorphisms $\zeta \to \zeta$ and $\zeta \to \zeta^{-1}$ in the Galois group of the cyclotomic extension $\mathbb{Q}(\zeta)/\mathbb{Q}$, which is Galois and of degree $\varphi(n)$ over $\mathbb{Q}$ (Theorem 58.12). So $\mathbb{Q}(\cos(2\pi/n))$ is an intermediate field of $\mathbb{Q}(\zeta)/\mathbb{Q}$ satisfying $|\mathbb{Q}(\zeta):\mathbb{Q}(\cos(2\pi/n))| = |\{\zeta,\zeta^{-1}\}| = 2$. Thus $|\mathbb{Q}(\cos(2\pi/n)):\mathbb{Q}| = \varphi(n)/2$. Hence, if $\cos(2\pi/n)$ is constructible, then $\varphi(n)/2$ and consequently also $\varphi(n)$ is a power of two. Let $n = 2^{a_0}p_1^{a_1}p_2^{a_2}\ldots p_r^{a_r}$ be the canonical decomposition of n into prime numbers, but possibly with $a_0 = 0$. Then

$$\varphi(n) = 2^{a_0-1}p_1^{a_1-1}p_2^{a_2-1}\ldots p_r^{a_r-1}(p_1 - 1)(p_2 - 1)\ldots(p_r - 1)$$

(the term 2^{a_0-1} is to be deleted in case $a_0 = 1$) and $\varphi(n)$ is a power of two if and only if $a_1 = a_2 = \cdots = a_r = 1$ and $p_i = 2^{k_i} + 1$ for some $k_i \in \mathbb{N}$. Here k_i cannot be divisible by an odd number t, for otherwise $2^t + 1$ would divide the prime number $2^{k_i} + 1$. Hence k_i is a power of two, say $k_i = 2^{m_i}$. Thus $p_i = 2^{2^{m_i}} + 1$. Prime numbers of the form $2^{2^m} + 1$ are called Fermat primes. It is easily verified that $2^{2^m} + 1$ is prime when $m = 0,1,2,3,4$, but $2^{2^5} + 1$ is not prime (it is divisible by 641). It is not known whether there are infinitely or finitely many Fermat primes. In fact, numbers $2^{2^m} + 1$ are known to be prime only in the cases $m = 0,1,2,3,4$. We obtain: *if a regular n-gon is constructible, then n has the form* $n = 2^{a_0}p_1p_2\ldots p_r$, *where* p_i *are distinct Fermat primes.* The converse of this statement is also true, and its proof is left to the reader.

Exercises

1. Let K be a field and define an extension E of K to be a radical extension of K if E is separable over K and there are elements $u_1, u_2, \ldots, u_n$ in E such that $E = K(u_1, u_2, \ldots, u_n)$ and one of the following is true:

(i) u_i is a root of a polynomial of the form $x^{h_i} - a_i$ over $K(u_1, \ldots, u_{i-1})$, where either $char\, K = 0$ or $char\, K = p \neq 0$ and p is relatively prime to h_i;

(ii) u_i is a root of polynomial of the form $x^p - x - a_i$ over $K(u_1, \ldots, u_{i-1})$, where $p = char\, K \neq 0$.

Prove that Theorem 59.8, Theorem 59.9, Theorem 59.10 and Theorem 59.11 remain valid with this new definition of radical extensions.

2. Let K be a field and define an extension E of K to be a radical extension of K if there are elements $u_1, u_2, \ldots, u_n$ in E such that $E = K(u_1, u_2, \ldots, u_n)$ and one of the following is true:

(i) u_i is a root of a polynomial of the form $x^{h_i} - a_i$ over $K(u_1, \ldots, u_{i-1})$,

(ii) u_i is a root of polynomial of the form $x^p - x - a_i$ over $K(u_1, \ldots, u_{i-1})$, where $p = char\, K \neq 0$.

Prove that Theorem 59.8 remains valid with this new definition of radical extensions if we assume E is normal over K. Discuss Theorem 59.9, Theorem 59.10 and Theorem 59.11.

3. Let K be a field, $f(x) \in K[x]$ an irreducible polynomial of degree $n \geqslant 5$, E a splitting field of $f(x)$ over K and r a root of $f(x)$ in E. Assume that $Aut_K E \cong S_5$. Prove the following assertions.

(i) $K(r)$ is not a Galois extension of K.

(ii) $|K(r):K| = n$.

(iii) $Aut_K K(r) \cong 1$.

(iv) If N is a normal closure of K over $K(r)$, then there is a subfield of N isomorphic to E.

(v) There is no radical extension R of K such that $K \subseteq K(r) \subseteq R$.

(This exercise shows the hypothesis that E be Galois over K is indespensable in Theorem 59.8.)

4. Prove that S_5 and A_5 are the only nonsolvable transitive subgroups of S_5. (Hint: Assume $\pi = (12345)$ is in such a subgroup G of S_5. There is a

transposition or a 3-cycle in G. In the first case, G contains all transpositions and $G = S_5$. In the second case, with η a 3-cycle in G, we have $\langle \pi^\eta \rangle \neq \langle \pi \rangle$ and there are 5^2 even permutations in G and $A_5 \leqslant G$.)

5. Let $f(x) \in \mathbb{Q}[x]$ be an irreducible polynomial of degree 5. Show that, if $f(x)$ has three real and two complex conjugate roots, then the Galois group $f(x)$ is S_5. (Hint: Use Ex. 1 and the discriminant.)

6. Find five irreducible polynomials in $\mathbb{Q}[x]$ whose Galois groups are S_5.

7. Show that $\sqrt[3]{2+\sqrt{-121}} + \sqrt[3]{2-\sqrt{-121}} = 4$ (Raffael Bombelli (ca. 1520-1572)).

8. Let K be a subfield of $\mathbb{R}$ and let $f(x)$ be a cubic polynomial in $K[x]$. Let D be the discriminant of $f(x)$. Prove that

(a) $D > 0$ if and only if $f(x)$ has three real distinct roots;

(b) $D < 0$ if and only if $f(x)$ has one real and two complex conjugate roots;

(c) $D = 0$ if and only if $f(x)$ has three real roots, one of which is repeated.

9. Let K be a subfield of $\mathbb{R}$ and let $f(x)$ be an irreducible polynomial in $K[x]$ such that $K(a)$ is a splitting field over K of $f(x)$, for any root a of $f(x)$. Show that there is no splitting field of $f(x)$ over K and a radical extension R of K satisfying $S \subseteq R \subseteq \mathbb{R}$.

10. Prove the converse of Theorem 59.32. (Hint: Show that, if the degree of K_n over $\mathbb{Q}$ is a power of two, so is the degree over $\mathbb{Q}$ of the normal closure of $\mathbb{Q}$ over K_n. Use Galois correspondence and Ex. 12 in §26.)

11. Prove that the angle 90° can and the angle 60° cannot be trisected by ruler and compass alone.

12. Show that, if n has the form $n = 2^{a_0} p_1 p_2 \cdots p_r$, where p_i are distinct Fermat primes, then a regular n-gon is constructible by ruler and compass alone.

Appendix

In this appendix, we want to point out some deficiencies of the naïve set theory we used in this book and discuss Zorn's lemma.

We used the term "set" informally to designate a collection of objects. G. Cantor (1845-1918), the founder of set theory, defines a set as "the collection of definite, well distinguished objects of our intuition or our thought to a whole." This definition is very general. The objects need not be mathematical objects. They can be concepts like charm, courage, miserability. In fact, Cantor writes in a letter that set theory belongs to metaphysics. But then, if there is no restriction on the objects forming a set, one is bound to admit "the set of all abstract concepts" or "the set of all sets". These sets are elements of themselves: the set of all abstract concepts is an abstract concept, and the set of all sets is a set. This general definition of a set leads now to logical paradoxes. Consider the set $M = \{A : A \notin A\}$ of all sets A such that $A \notin A$. (This is not unreasonable, for example the set of all girls is not a girl.) Now $M \in M$ implies, by the definition of M, that $M \notin M$ and $M \notin M$ implies $M \in M$. So $M \in M$ if and only if $M \notin M$, a contradiction.

This paradox is known as Russell's Paradox (B. Russell, 1872-1970). It was also known to G. Cantor, to D. Hilbert and to E. Zermelo (1871-1953). (A similar parodox is due to C. Burali-Forti (1861-1931): the set of all ordinal numbers is an ordinal number, hence is smaller than itself.)

These and analogous paradoxes indicate the necessity of a sounder approach to set theory. In the Gödel-Bernay axiomatic set theory, there are three primitive undefined concepts: class, membership and equality. A class is intuitively a collection of objects, to which we referred to as a set up to now (and what we called subset is a *subclass* in axiomatic set theory). The term *set* is reserved for the special type of classes A for which a class B exist such that $A \in B$. A class which is not a set is called a *proper class*. Loosely speaking, a set is a "small" class and a proper class is "wildly large." Axioms of set theory ensure that the subclass, union, intersection, difference and cartesian product of sets are sets. The class of all subclasses of a set is also a set.

Russell's paradox does not arise in the Gödel-Bernay formalism: the class $\{A$ is a set and $A \notin A\}$ is a proper class. (The class of all ordinal numbers is likewise a proper class.) We have the class of all sets, the class of all groups, the class of all rings, the class of all rings with identity, the class of all K-vector spaces, the class of all fields, the class of all abelian groups, the class of all solvable groups etc. These are not sets.

Thus isomorphism of groups is an equivalence relation on the *class* of all groups, not on the *set* of all groups. In like manner, isomorphism of rings is an equivalence relation on the *class* of all rings and isomorphism of K-vector spaces is an equivalence relation on the *class* of all K-vector spaces.

Aside from the classes of sets, groups, rings, K-vector spaces, solvable groups, all the classes which we called *set* in the text are indeed sets (not proper classes) in axiomatic set theory.

*

* *

Now Zorn's Lemma. Zorn's Lemma states that any partially ordered set, in which every totally ordered subset has an upper bound, possesses a maximal element. We explain the new terms in its formulation.

A1 Definition: Let S be a nonempty set and R a relation on S. If R satisfies the following conditions, then R is called a *partial order on* S:

(i) $a\ R\ a$ for all $a \in S$ (R is reflexive),

(ii) for all $a,b \in S$, if $a\ R\ b$ and $b\ R\ a$, then $a = b$ (R is antisymmetric),

(iii) for all $a,b,c \in S$, if $a\ R\ b$ and $b\ R\ c$, then $a\ R\ c$ (R is transitive).

The set S is then said to be *partially ordered by* R and S is called a *partially ordered set*.

We shall almost never use the symbol "R" to designate a partial order. We use "$\leqslant$", "$\subseteq$" or a similar symbol for this purpose. Then the conditions above assume the more palatable form

(i) $a \leqslant a$ for all $a \in S$,

(ii) $a \leqslant b$ and $b \leqslant a \implies a = b$ (for all $a,b \in S$),

(iii) $a \leqslant b$ and $b \leqslant c \implies a \leqslant c$ (for all $a,b,c \in S$).

It is worthwhile to compare partial order relations with equivalence relations. Both relations are reflexive and transitive. They differ in that partial orders are antisymmetric and equivalence relations are symmetric. These are quite opposite conditions and partial orders are indeed very different from equivalence relations.

The reader shoul be careful about the antisymmetry property. It does *not* say

$$a \leqslant b \implies b \not\leqslant a \qquad (a,b \in S)$$

(where $b \not\leqslant a$ is the negation of $b \leqslant a$).

A2 Examples: (a) Consider the set $\mathbb{N}$ of natural numbers and the usual order relation $\leqslant$. Hence $a \leqslant b$ means $b - a \in \mathbb{N}$ or $a = b$. It is easy to see that $\leqslant$ is a partial order on $\mathbb{N}$.

(b) The usual order relation $\leqslant$ (for which $a \leqslant b$ means $b - a \in \mathbb{N}$ or $a = b$) is a partial order on $\mathbb{Z}$.

(c) The sets $\mathbb{Q}$ and $\mathbb{R}$ are partially ordered by $\leqslant$.

(d) Divisibility is a partial order on $\mathbb{N}$, because

(i) $a|a$,

(ii) $a|b$ and $b|a \implies a = b$,

(iii) $a|b$ and $b|c \implies a|c$

for all $a,b,c \in \mathbb{N}$.

(e) Divisibility is not a partial order on $\mathbb{Z}$, because it is not a relation on $\mathbb{Z}$: $a|b$ is meaningful only when $a \neq 0$.

(f) The "is strictly less than" relation $<$ on $\mathbb{Z}$ is not a partial order on $\mathbb{Z}$ because it is not reflexive. Is it antisymmetric?

(g) Let $\mathfrak{A}$ be the set of all subsets of a given set A. Then $\mathfrak{A}$ is partially ordered by the set inclusion $\subseteq$. Indeed, for any subsets B,C,D of A, we have

(i) $B \subseteq B$,

(ii) $B \subseteq C$ and $C \subseteq B$ implies $B = C$,

(iii) $B \subseteq C$ and $C \subseteq D$ implies $B \subseteq D$.

On the other hand, $\mathfrak{A}$ is not partially ordered by the proper set inclusion $\subset$ since $\subset$ is not reflexive.

(h) Let $\mathfrak{A}$ be the set of all subgroups of a given group G. Then $\mathfrak{A}$ is partially ordered by the "is a subgroup of" relation $\leqslant$. Indeed, for any subgroups B,C,D of G, we have

(i) $H \leqslant H$,

(ii) $H \leqslant J$ and $J \leqslant H$ implies $H = J$,

(iii) $H \leqslant J$ and $J \leqslant K$ implies $H \leqslant K$.

On the other hand, $\mathfrak{A}$ is not partially ordered by the "is a proper subgroup of" relation $<$ since $<$ is not reflexive.

(i) Let $\mathfrak{A}$ be the set of all subgroups of a given group G. Then $\mathfrak{A}$ is not a partially ordered by the "is a normal subgroup of" relation $\trianglelefteq$ because $\trianglelefteq$ is not transitive (cf. Example 18.5(i)).

(j) Let $\mathfrak{A}$ be the set of all ideals of a given ring R. Then $\mathfrak{A}$ is partially ordered by the set inclusion $\subseteq$.

(k) Let K be a field and V a vector space over K. Let $\mathfrak{A}$ be the set of all K-linearly independent subsets of V. Then $\mathfrak{A}$ is partially ordered by the set inclusion $\subseteq$.

The natural numbers $\mathbb{N}$ is partially ordered by $\leqslant$ as well as by $|$. We know that, for any two natural numbers n and m, at least one of $n \leqslant m$ and $m \leqslant n$ is true, whereas neither $n|m$ nor $m|n$ ought to hold. Thus any two natural numbers can be compared by the partial order $\leqslant$, but not by the order $|$.

A3 **Definition:** Let S be a nonempty set and $\leqslant$ a partial order on S. Let a,b be elements of S. If $a \leqslant b$ or $b \leqslant a$, then a is said to be *comparable to* b (in this case, b is also comparable to a and we also say then that a and b are comparable). If any two elements of S are comparable, then $\leqslant$ is called a *total order on* S. The set S is then said to be *totally ordered by* $\leqslant$ and S is called a *totally ordered set*.

Thus $\mathbb{N}$ is totally ordered by $\leqslant$ but not by $|$. Here are some more examples.

A4 Examples: (a) $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ are totally ordered by $\leqslant$.

(b) The set $\{\varnothing, \{1\}, \{1,2\}, \{1,2,3\}\}$ is totally ordered by $\subseteq$. The set $\{\varnothing, \{1\}, \{2\}, \{3\}, \{1,2\}\}$ is not totally ordered by $\subseteq$.

(c) The set $\{1, \langle\rho^2\rangle, \langle\rho\rangle\}$ of subgroups of $D_8 = \langle\rho, \sigma: \rho^4 = \sigma^2 = 1\rangle$ is totally ordered by the "is a subgroup of" relation $\leqslant$. The set $\{1, \langle\sigma\rangle, \langle\rho^2,\sigma\rangle\}$ of subgroups of D_8 is also totally ordered by $\leqslant$. The set $\{1, \langle\sigma\rangle, \langle\rho\rangle\}$ of subgroups of D_8 is not totally ordered by $\leqslant$.
Can you describe the groups such that the set of all subgroups is totally ordered by $\leqslant$?

(d) The finite set $\{1,2,4,8,16\}$ and the infinite set $\{1,2,4,8,16,32,64,\ldots\}$ are totally ordered subsets of $\mathbb{N}$ when $\mathbb{N}$ is ordered by $|$.

(e) Any nonzero rational number can be written in the form $7^a\frac{u}{v}$, where $a,u,v \in \mathbb{Z}$ and $7\nmid u$, $7\nmid v$. For any two rational numbers $x = 7^a\frac{u}{v}$, $y = 7^b\frac{s}{t}$, where $7\nmid uvst$, we write $x \ll y$ provided $a \leqslant b$. Then $\mathbb{Q}\backslash\{0\}$ is totally ordered by $\ll$.

A5 Definition: Let S be a nonempty set and $\leqslant$ a partial order on S. An element m of S is called a *maximal element of* S if the following implication holds:

$$x \in S \text{ and } m \leqslant x \quad \Longrightarrow \quad m = x.$$

A6 Examples: (a) Let G be a nontrivial group and $\mathcal{S}$ be the set of all proper normal subgroups of G, partially ordered by the relation $\subseteq$. A maximal normal subgroup of G, in the sense of Definition 27.5, is a maximal element of $\mathcal{S}$, and conversely.

(b) Let R be a ring distinct from the null ring, and $\mathcal{S}$ be the set of all proper ideals of R, partially ordered by the relation $\subseteq$. A maximal

element of $\mathfrak{S}$ is called a maximal ideal of G. Thus M is a maximal ideal of R if and only if there is no ideal A of R such that $M \subset A \subset R$.

(c) Let V be a vector space over a field K. Let $\mathfrak{S}$ be the set of all K-linearly independent subsets of V, partially ordered by $\subseteq$. The maximal elements of $\mathfrak{S}$ are exactly the K-bases of V. This assertion will be proved below (Theorem A 10).

(d) $\mathbb{Z}$, when partially ordered by $\leqslant$, does not have any maximal element.

(e) Let $S = \{a,b,c,d,e,f\}$ and define a relation $\leqslant$ on S by setting

$$\begin{array}{lllll} a \leqslant a, & b \leqslant b, & c \leqslant c, & d \leqslant d; & e \leqslant e, \quad f \leqslant f, \\ a \leqslant d, & a \leqslant d, & b \leqslant c, & c \leqslant f, & d \leqslant f. \end{array}$$

Then S is partially ordered by $\leqslant$. We see that e and f are maximal elements of S. Can you find the totally ordered subsets of S? There are thirteen of them.

A partially ordered set need not have a maximal element. If it does have one, a maximal element need not be unique, i.e., the set may have more than one maximal elements.

A maximal element m of a partially ordered set S is not necessarily "bigger" than all the other elements. More precisely, it need not be the case that $a \leqslant m$ for all $a \in S$. This condition is stronger than maximality. A maximal element m is "bigger" than all the other elements *that are comparable to* m.

A7 Definition: Let S be a nonempty set and $\leqslant$ a partial order on S. Let A be a nonempty subset of S. An element b of S is called an *upper bound of* A if $a \leqslant b$ for *all* $a \in A$.

It is implicit in this definition that b is comparable to all the elements of A. An upper bound of A is not necessarily an element of A. A nonempty subset of a partially ordered set need not have an upper bound.

A8 Examples: (a) $\mathbb{N}$ is a nonempty subset of $\mathbb{Z}$, which is partially (in fact totally) ordered by $\leqslant$ and an upper bound of $\mathbb{N}$ does not exist.

(b) $\mathbb{Q}$ is partially ordered by $\leqslant$. Consider $A = \{a \in \mathbb{Q}: a^2 < 2\}$. There are many upper bounds of A, for instance 3/2, 2, 10, 10^{10}. Is $\sqrt{2}$ an uper bound of A?

(c) Let $\mathfrak{S}$ be the set of all subsets of a given set T, partially ordered by $\subseteq$. A subset $\mathfrak{L} = \{T_1, T_2, \ldots, T_n\}$ of $\mathfrak{S}$ has an upper bound in $\mathfrak{S}$, for instance $\bigcup_{i=1}^{n} T_i$ is an upper bound of $\mathfrak{L}$.

(d) Let $\mathfrak{S}$ be the set of all proper subsets of $\{a,b,c\}$, partially ordered by $\subseteq$. The subset $\{\{a\}, \{b,c\}\}$ of $\mathfrak{S}$ does not have any upper bound in $\mathfrak{S}$.

Let us recall Zorn's lemma.

A9 Zorn's Lemma: *Let S be a partially ordered set. If every totally ordered subset of S contains an upper bound, then S has a maximal element.*

Some explanations are now in order. We make use of Zorn's Lemma to prove that certain things exist. The idea is to consider the object whose existence we want to show as a maximal element in a partially ordered set S. The set S and the partial order $\leqslant$ is dictated by the nature of the problem at hand. The partial order is usually a natural order like set inclusion or the "is a subgroup of" relation. Since we speak of partial order, the set S must be nonempty. In practice, it is either the easiest or the hardest part to prove that S is nonempty. After establishing that S is not empty, we introduce a partial order $\leqslant$ on S such that the object we want to show to exist is a maximal element of S. Then we take an arbitrary totally ordered subset A of S and investigate if A contains an upper bound. Here it is important to note that the upper bound is required to be an element of A. Thus we must find an upper bound u of

A and also ascertain that u is not merely in S, but in fact in A. If we can show that *any* totally ordered subset A of S does contain an upper bound *in* A, we conclude, by Zorn's Lemma, that S has a maximal element. The existence of a maximal element of S is thereby proved.

As an illustration, we prove an important theorem: any vector space has a basis.

A10 Theorem: *Let V be a vector space over a field K. Then V has a K-basis.*

Proof: If $V = 0$, then V has a K-basis, namely $\varnothing$, and the theorem is proved in this trivial case. We assume now $V \neq 0$. Let $\mathfrak{A}$ be the set of all K-linearly independent subsets of V. As $V \neq 0$, there is a nonzero vector u in V and $\{u\}$ is a linearly independent subset of V. Thus $\{u\} \in \mathfrak{A}$ and $\mathfrak{A}$ is not empty. $\mathfrak{A}$ is partially ordered by $\subseteq$. We prove that any maximal element of $\mathfrak{A}$ is a K-basis of V. To prove this assertion, we must show that, if $B \in \mathfrak{A}$ is a maximal element of $\mathfrak{A}$, then B spans V over K.

If $B \in \mathfrak{A}$ is a maximal element of $\mathfrak{A}$ and $s_K(B) \neq V$, then there is a $v \in V$ with $v \notin s_K(B)$. Let $A = B \cup \{v\}$. Since $v \notin s_K(B)$, we have $v \notin B$ and therefore $B \subset A$. Here A cannot be an element of $\mathfrak{A}$ because $B \subset A$ and B is a maximal element of $\mathfrak{A}$. Thus A is a K-linearly dependent subset of V. This means that A has a *finite* K-linearly dependent subset C. Here $C \subseteq B$ is impossible because B is a K-linearly independent subset of V. Hence we necessarily have $v \in C$. Let $C = \{v, v_1, v_2, \ldots, v_n\}$. Here $\{v_1, v_2, \ldots, v_n\} \subseteq B$. As C is linearly dependent over K, there are scalars $\alpha, \alpha_1, \alpha_2, \ldots, \alpha_n$, not all of them zero, such that

$$\alpha v + \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0.$$

Now $\alpha = 0$ or $\alpha \neq 0$. In the first case, we have $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$, with $\{v_1, v_2, \ldots, v_n\} \subseteq B$ and not all of $\alpha_1, \alpha_2, \ldots, \alpha_n$ being zero, and this means $\{v_1, v_2, \ldots, v_n\}$ is K-linearly dependent, so B is also K-linearly dependent, contrary to $B \in \mathfrak{A}$. In the second case, there is an inverse α^{-1} of α in K, and we get

$$v = (-\alpha^{-1}\alpha_1)v_1 + (-\alpha^{-1}\alpha_2)v_2 + \cdots + (-\alpha^{-1}\alpha_n)v_n \in s_K(B),$$

contrary to $v \notin s_K(B)$.

Thus the assumption $s_K(B) \neq V$ leads to contradictions. This forces $s_K(B) = V$ and B is a K-basis of V, as was to be proved.

We now employ Zorn's Lemma to conclude that $\mathfrak{A}$ has a maximal element. This will prove that V has a K-basis.

We have to show that any totally ordered subset of $\mathfrak{A}$ has an upper bound in $\mathfrak{A}$. Let $C = \{A_i : i \in I\}$ be a totally ordered subset of $\mathfrak{A}$. It is natural to expect that $E := \bigcup_{i \in I} A_i$ is an upper bound of C. We certainly have $A_j \subseteq E$ for any $A_j \in C$ but we must also check that $E \in \mathfrak{A}$. To this end, we must prove that E is K-linearly independent, that is to say, that every finite subset of E is K-linearly independent.

Let $\{v_1, v_2, \ldots, v_n\}$ be finite subset of E. We have $v_k \in A_{i_k}$ for some suitable $A_{i_k} \in C$. Since C is totally ordered, either $A_{i_1} \subseteq A_{i_2}$ or $A_{i_2} \subseteq A_{i_1}$. Let us put $U_2 = A_{i_2}$ if $A_{i_1} \subseteq A_{i_2}$ and $U_2 = A_{i_1}$ if $A_{i_2} \subseteq A_{i_1}$ so that $\{v_1, v_2\} \subseteq U_2$. Now either $U_2 \subseteq A_{i_3}$ or $A_{i_3} \subseteq U_2$. Let us put $U_3 = A_{i_3}$ if $U_2 \subseteq A_{i_3}$ and $U_3 = U_2$ if $A_{i_3} \subseteq U_2$ so that $\{v_1, v_2, v_3\} \subseteq U_3$. Now either $U_3 \subseteq A_{i_4}$ or $A_{i_4} \subseteq U_3$. Let us put $U_4 = A_{i_4}$ if $U_3 \subseteq A_{i_4}$ and $U_4 = U_3$ if $A_{i_4} \subseteq U_3$ so that $\{v_1, v_2, v_3, v_4\} \subseteq U_4$. Proceeding in this way, we see that $\{v_1, v_2, \ldots, v_n\} \subseteq U_n$, where U_n is one of $A_{i_1}, A_{i_2}, \ldots, A_{i_n}$. From $U_n \in C \subseteq \mathfrak{A}$, we infer that U_n is K-linearly independent, and from $\{v_1, v_2, \ldots, v_n\} \subseteq U_n$, we infer that $\{v_1, v_2, \ldots, v_n\}$ is K-linearly independent. Thus every finite subset of E is K-linearly independent and so $E \in \mathfrak{A}$.

This shows that every totally ordered subset of $\mathfrak{A}$ has an upper bound in $\mathfrak{A}$. By Zorn's Lemma, there is a maximal element B of $\mathfrak{A}$. So there is a K-basis B of V. □

Redefining $\mathfrak{A}$ in the preceding proof to be the set of all K-linearly independent subsets of V *containing* T, and using $T \in \mathfrak{A}$ to show that $\mathfrak{A}$ is not empty, we get the following generalization of Theorem 42.14.

A11 Theorem: *Let V be a vector space over a field K and let T be a K-linearly independent subset of V. Then there is a K-basis B of V such that $T \subseteq B$.* □

The theorem that any field K has an algebraic closure can also be proved by Zorn's lemma. One is tempted to consider here the set (set! set? *Set*!?) of all algebraic extensions of K and find a maximal algebraic extensions of K. But the problem is that the class of all algebraic extensions of K is not a set, and Zorn's lemma cannot be used so uncritically. One must first find a *set* of algebraic extensions of K to apply Zorn's lemma. This is done by showing that an algebraic extension of K has a cardinal number not much greater than the cardinality of K, and algebraic extensions of K can be assumed to be *subsets* of a set M containing K, provided the cardinality of M is large enough. For details, the reader is referred to the literature.

Zorn's lemma is equivalent to an axiom of set theory, known as the axiom of choice. The axioms of set theory imply Zorn's lemma. Likewise, using the axioms of set theory except for the axiom of choice, and admitting Zorn's lemma as an axiom, one can prove the axiom of choice. The axiom of choice is independent of the remaining axioms of set theory. We refer the reader to the literature for more information about this topic.

References

Albert A. A. *Fundamental Concepts of Higher Algebra*, 1966, The University of Chicago Press, Chicago-London.

Artin Emil, *Galoissche Theorie*, 1973, Harrie Deutsch Verlag, Zürich-Frankfurt a. M.

Boyer Carl B., *A History of Mathematics*, 1968, John Wiley and Sons, Inc., New York-London-Sydney.

Carmichael R. D., *Groups of Finite Order*, 1956, Dover Publications Inc., New York.

Dean R. A., *Elements of Abstract Algebra*, 1966, John Wiley and Sons, Inc., New York-London-Sydney.

Ebbinghaus H.-D., Hermes. H., Hirzebruch F., Koecher M., Mainzer K., Prestel A., Remmert R., *Zahlen*, 1983, Grundwissen Mathematik 1, Springer Verlag, Berlin-Heidelberg-New York.

Edwards Harold M., *Galois Theory*, 1984, Graduate Texts in Mathematics 101, Springer Verlag, Berlin-Heidelberg-New York-Tokyo.

Gaal L., *Classical Galois Theory*, 1979, Chelsea Publishing Company.

Hasse Helmut, Höhere *Algebra* I, 1969, Sammlung Göschen Band 931, Walter de Greuter & Co, Berlin.

Hasse Helmut, Höhere *Algebra* II, 1967, Sammlung Göschen Band 932, Walter de Greuter & Co, Berlin.

Hecke Erich, *Vorlesungen über die Theorie der Algebraischen Zahlen*, Reprint 1970, Chelsea Publishing Company, New York.

Herstein I. N., *Topics in Algebra*, 1975, John Wiley & Sons, Inc., New York-London-Sydney-Toronto.

Hungerford Thomas W., *Algebra*, 1974, Graduate Texts in Mathematics 73, Springer Verlag, Berlin-Heidelberg-New York.

Huppert B., *Endliche Gruppen* I, 1979, Grundlehren der Mathematische Wissenschaften 134, Springer Verlag, Berlin-Heidelberg-New York.

Hölder Otto, "Galoissche Theorie mit Anwendungen," *Encyklopädie der Mathematischen Wissanschaften und ihre Grenzgebiete*, I: *Arithmetik und Algebra* IB3c,d, 1898-1904, Teubner Verlag, Leipzig.

Ireland K., Rosen M., *A Classical Introduction to Modern Number Theory*, 1980, Graduate Texts in Mathematics 84, Springer Verlag, Berlin-Heidelberg-New York.

Kaplansky Irving, *Fields and Rings*, 1972, The University of Chicago Press, Chicago and London.

Kiernan B. Melvin, "The Development of Galois Theory from Lagrange to Artin," *Arch. Hist. Exact. Sci.*, 8 (1971) 40-154.

Klein Felix, *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*, Reprint 1979, Springer Verlag, Berlin-Heidelberg-New York.

Koch Helmut, *Einführung in die Klassische Mathematik*, 1986, Springer Verlag, Berlin-Heidelberg-New York-London-Paris-Tokyo.

Konforowitsch A. G., *Guten Tag, Herr Archimedes*, 1986, Deutsch Taschenbücher Band 50, Harrie Deutsch Verlag, Thun-Frankfurt a. M.

Landau Edmund, *Vorlesungen über Zahlentheorie* (3 Bd), Reprint 1969, Chelsea Publishing Company, New York.

Lang Serge, *Algebra*, 1969, Addison-Wesley Publishing Company Inc., Reading, Massachusetts.

Lang Serge, *Linear Algebra*, 1978, Addison-Wesley Publishing Company Inc., Reading, Massachusetts.

Lowenthal Franklin, *Linear Algebra with Differential Equations*, 1975, John Wiley & Sons, Inc., New York-London-Sydney-Toronto.

Martin G. E., *Transformation Geometry*, 1982, Undergraduate Texts in Mathematics, Springer Verlag, Berlin-Heidelberg-New York.

Meschkowski H., *Einführung in die Moderne Mathematik*, 1964, Bibliographisches Institut, Hochschultaschenbücher Verlag, Mannheim.

Pierce R. S., *Associative Algebras*, 1980, Graduate Texts in Mathematics 88, Springer Verlag, Berlin-Heidelberg-New York.

Rose S. John, *A Course on Group Theory*, 1978, Cambridge University Press, Cambridge-London-New York-Melbourne.

Smith D. E., *History of Mathematics* I, 1958, Dover Publications Inc., New York.

van der Waerden B. L., *Algebra* I, 1971, Heidelberger Taschenbücher 12, Springer Verlag, Berlin-Heidelberg-New York.

Weber Heinrich, *Lehrbuch der Algebra* I *und* II, Reprint 1979, Chelsea Publishing Company, New York.

Winter D. J., *The Structure of Fields*, 1974, Graduate Texts in Mathematics 16, Springer Verlag, Berlin-Heidelberg-New York.

Zariski O., Samuel P., *Commutative Algebra* I, 1975, Graduate Texts in Mathematics 28, Springer Verlag, Berlin-Heidelberg-New York.