6 characteristic polynomial must be a 6 × 6 matrix. The polynomial $(x^4 - 1)(x^2 - 1)$ factors into irreducibles in $\mathbb{Q}[x]$ as $(x - 1)^2(x + 1)^2(x^2 + 1)$. Since the minimal polynomial $m_A(x)$ for A has the same roots as $c_A(x)$ it follows that $(x-1)(x+1)(x^2+1)$ divides $m_A(x)$. Suppose $a_1(x), \dots, a_m(x)$ are the invariant factors of some A, so $a_m(x) = m_A(x)$, $a_i(x) \mid a_{i+1}(x)$ (in particular, all the invariant factors divide $m_A(x)$) and $a_1(x)a_2(x)\cdots a_m(x) = (x^4 - 1)(x^2 - 1)$. One easily sees that the only permissible lists under these constraints are

(a) $(x - 1)(x + 1), \quad (x - 1)(x + 1)(x^2 + 1)$
(b) $x - 1, \quad (x - 1)(x + 1)^2(x^2 + 1)$
(c) $x + 1, \quad (x - 1)^2(x + 1)(x^2 + 1)$
(d) $(x - 1)^2(x + 1)^2(x^2 + 1)$.

One can now easily write out the corresponding direct sums of companion matrices to obtain representatives of the 4 similarity classes. We shall see in the next section that there are still only 4 similarity classes even in $M_6(\mathbb{C})$.

(5) In this example we find all similarity classes of 3 × 3 matrices A with entries from $\mathbb{Q}$ satisfying $A^6 = I$. For each such A, its minimal polynomial divides $x^6 - 1$ and in $\mathbb{Q}[x]$ the complete factorization of this polynomial is

$$x^6 - 1 = (x - 1)(x + 1)(x^2 - x + 1)(x^2 + x + 1).$$

Conversely, if B is any 3 × 3 matrix whose minimal polynomial divides $x^6 - 1$, then $B^6 = I$. The only restriction on the minimal polynomial for B is that its degree is at most 3 (by the Cayley–Hamilton Theorem). The only possibilities for the minimal polynomial of such a matrix A are therefore

(a) $x - 1$
(b) $x + 1$
(c) $x^2 - x + 1$
(d) $x^2 + x + 1$
(e) $(x - 1)(x + 1)$
(f) $(x - 1)(x^2 - x + 1)$
(g) $(x - 1)(x^2 + x + 1)$
(h) $(x + 1)(x^2 - x + 1)$
(i) $(x + 1)(x^2 + x + 1)$.

Under the constraints of the rational canonical form these give rise to the following permissible lists of invariant factors:

(i) $x - 1, \quad x - 1, \quad x - 1$
(ii) $x + 1, \quad x + 1, \quad x + 1$
(iii) $x - 1, \quad (x - 1)(x + 1)$
(iv) $x + 1, \quad (x - 1)(x + 1)$
(v) $(x - 1)(x^2 - x + 1)$
(vi) $(x - 1)(x^2 + x + 1)$
(vii) $(x + 1)(x^2 - x + 1)$
(viii) $(x + 1)(x^2 + x + 1)$.

Note that it is impossible to have a suitable set of invariant factors if the minimal polynomial is $x^2 + x + 1$ or $x^2 - x + 1$. One can now write out the corresponding rational canonical forms; for example, (i) is I, (ii) is $-I$, and (iii) is

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

Note also that another way of phrasing this result is that any 3 × 3 matrix with entries from $\mathbb{Q}$ whose order (multiplicatively, of course) divides 6 is similar to one of these 8 matrices, so this example determines all elements of orders 1,2,3 and 6 in the group $GL_3(\mathbb{Q})$ (up to similarity).

EXERCISES

1. Prove that similar linear transformations of V (or $n \times n$ matrices) have the same characteristic and the same minimal polynomial.
2. Let M be as in Lemma 19. Prove that the minimal polynomial of M is the least common multiple of the minimal polynomials of $A_1, \dots, A_k$.
3. Prove that two 2×2 matrices over F which are not scalar matrices are similar if and only if they have the same characteristic polynomial.
4. Prove that two 3×3 matrices are similar if and only if they have the same characteristic and same minimal polynomials. Give an explicit counterexample to this assertion for 4×4 matrices.
5. Prove directly from the fact that the collection of *all* linear transformations of an n dimensional vector space V over F to itself form a vector space over F of dimension n^2 that the minimal polynomial of a linear transformation T has degree at most n^2.
6. Prove that the constant term in the characteristic polynomial of the $n \times n$ matrix A is $(-1)^n \det A$ and that the coefficient of x^{n-1} is the negative of the sum of the diagonal entries of A (the sum of the diagonal entries of A is called the *trace* of A). Prove that $\det A$ is the product of the eigenvalues of A and that the trace of A is the sum of the eigenvalues of A.
7. Determine the eigenvalues of the matrix

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

8. Verify that the characteristic polynomial of the companion matrix

$$\begin{pmatrix} 0 & 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & 0 & \dots & 0 & -a_1 \\ 0 & 1 & 0 & \dots & 0 & -a_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -a_{n-1} \end{pmatrix}$$

is

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

9. Find the rational canonical forms of

$$\begin{pmatrix} 0 & -1 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} c & 0 & -1 \\ 0 & c & 1 \\ -1 & 1 & c \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 422 & 465 & 15 & -30 \\ -420 & -463 & -15 & 30 \\ 840 & 930 & 32 & -60 \\ -140 & -155 & -5 & 12 \end{pmatrix}.$$

10. Find all similarity classes of 6×6 matrices over $\mathbb{Q}$ with minimal polynomial $(x+2)^2(x-1)$ (it suffices to give all lists of invariant factors and write out some of their corresponding matrices).

11. Find all similarity classes of 6×6 matrices over $\mathbb{C}$ with characteristic polynomial $(x^4-1)(x^2-1)$.

12. Find all similarity classes of 3×3 matrices A over $\mathbb{F}_2$ satisfying $A^6 = I$ (compare with the answer we computed over $\mathbb{Q}$). Do the same for 4×4 matrices B satisfying $B^{20} = I$.

13. Prove that the number of similarity classes of 3×3 matrices over $\mathbb{Q}$ with a given characteristic polynomial in $\mathbb{Q}[x]$ is the same as the number of similarity classes over any extension field of $\mathbb{Q}$. Give an example to show that this is not true in general for 4×4 matrices.

14. Determine all possible rational canonical forms for a linear transformation with characteristic polynomial $x^2(x^2+1)^2$.

15. Determine up to similarity all 2×2 rational matrices (i.e., $\in M_2(\mathbb{Q})$) of precise order 4 (multiplicatively, of course). Do the same if the matrix has entries from $\mathbb{C}$.

16. Show that $x^5 - 1 = (x-1)(x^2-4x+1)(x^2+5x+1)$ in $\mathbb{F}_{19}[x]$. Use this to determine up to similarity all 2×2 matrices with entries from $\mathbb{F}_{19}$ of (multiplicative) order 5.

17. Determine representatives for the conjugacy classes for $GL_3(\mathbb{F}_2)$. [Compare your answer with Theorem 15 and Proposition 14 of Chapter 6.]

18. Let V be a finite dimensional vector space over $\mathbb{Q}$ and suppose T is a nonsingular linear transformation of V such that $T^{-1} = T^2 + T$. Prove that the dimension of V is divisible by 3. If the dimension of V is precisely 3 prove that all such transformations T are similar.

19. Let V be the infinite dimensional real vector space

$$\mathbb{R}^\infty = \{(a_0, a_1, a_2, \ldots) \mid a_0, a_1, a_2, \cdots \in \mathbb{R}\}.$$

Define the map $T : V \to V$ by $T(a_0, a_1, a_2, \ldots) = (0, a_0, a_1, a_2, \ldots)$. Prove that T has no eigenvectors.

20. Let ℓ be a prime and let $\Phi_\ell(x) = \frac{x^\ell - 1}{x-1} = x^{\ell-1} + x^{\ell-2} + \ldots + x + 1 \in \mathbb{Z}[x]$ be the ℓ^{th} cyclotomic polynomial, which is irreducible over $\mathbb{Q}$ (Example 4 following Corollary 9.14). This exercise determines the smallest degree of a factor of $\Phi_\ell(x)$ modulo p for any prime p and so in particular determines when $\Phi_\ell(x)$ is irreducible modulo p. (This actually determines the complete factorization of $\Phi_\ell(x)$ modulo p — cf. Exercise 8 of Section 13.6.)

(a) Show that if $p = \ell$ then $\Phi_\ell(x)$ is divisible by $x - 1$ in $\mathbb{F}_\ell[x]$.

(b) Suppose $p \neq \ell$ and let f denote the order of p in $\mathbb{F}_\ell^\times$, i.e., f is the smallest power of p with $p^f \equiv 1 \bmod \ell$. Show that $m = f$ is the first value of m for which the group $GL_m(\mathbb{F}_p)$ contains an element A of order ℓ. [Use the formula for the order of this group at the end of Section 11.1.]

(c) Show that $\Phi_\ell(x)$ is not divisible by any polynomial of degree smaller than f in $\mathbb{F}_p[x]$ [consider the companion matrix for such a divisor and use (b)]. Let $m_A(x) \in \mathbb{F}_p[x]$ denote the minimal polynomial for the matrix A in (b) and conclude that $m_A(x)$ is irreducible of degree f and divides $\Phi_\ell(x)$ in $\mathbb{F}_p[x]$.

(d) In particular, prove that $\Phi_\ell(x)$ is irreducible modulo p if and only if $l-1$ is the smallest power of p which is congruent to 1 modulo ℓ, i.e., p is a primitive root modulo ℓ.

21. Prove that the first two elementary row and column operations described before Theorem 21 do not change the determinant of the matrix and the third elementary operation multiplies the determinant by a unit. Conclude from Theorem 21 that the characteristic polynomial of A differs by a unit from the product of the invariant factors of A. Since both these polynomials are monic by definition, conclude that they are equal (this gives an alternate proof of Proposition 20).

The following exercises outline the proof of Theorem 21. They carry out explicitly the construction described in Exercises 16 to 19 of the previous section for the Euclidean Domain $F[x]$. Let V be an n-dimensional vector space with basis $v_1, v_2, \dots, v_n$ and let T be the linear transformation of V defined by the matrix A and this choice of basis, i.e., T is the linear transformation with

$$T(v_j) = \sum_{i=1}^{n} a_{ij} v_i, \qquad j = 1, 2, \dots, n$$

where $A = (a_{ij})$. Let $F[x]^n$ be the free module of rank n over $F[x]$ and let $\xi_1, \xi_2, \dots, \xi_n$ denote a basis. Then we have a natural surjective $F[x]$-module homomorphism

$$\varphi : F[x]^n \to V$$

defined by mapping ξ_i to v_i, $i = 1, 2, \dots, n$. As indicated in the exercises of the previous section the invariant factors for the $F[x]$-module V can be determined once we have determined a set of generators and the corresponding relations matrix for $\ker\varphi$. Since by definition x acts on V by the linear transformation T, we have

$$x(v_j) = \sum_{i=1}^{n} a_{ij} v_i, \quad j = 1, 2, \dots, n.$$

22. Show that the elements

$$\nu_j = -a_{1j}\xi_1 - \cdots - a_{j-1\,j}\xi_{j-1} + (x - a_{jj})\xi_j - a_{j+1\,j}\xi_{j+1} - \cdots - a_{nj}\xi_n$$

for $j = 1, 2, \dots, n$ are elements of the kernel of φ.

23. (a) Show that $x\xi_j = \nu_j + f_j$ where $f_j \in F\xi_1 + \cdots + F\xi_n$ is an element in the F-vector space spanned by $\xi_1, \dots, \xi_n$.

(b) Show that

$$F[x]\xi_1 + \cdots + F[x]\xi_n = (F[x]\nu_1 + \cdots + F[x]\nu_n) + (F\xi_1 + \cdots + F\xi_n).$$

24. Show that $\nu_1, \nu_2, \dots, \nu_n$ generate the kernel of φ. [Use the previous result to show that any element of $\ker\varphi$ is the sum of an element in the module generated by $\nu_1, \nu_2, \dots, \nu_n$ and an element of the form $b_1\xi_1 + \cdots + b_n\xi_n$ where the b_i are elements of F. Then show that such an element is in $\ker\varphi$ if and only if all the b_i are 0 since $v_1, \dots, v_n$ are a basis for V over F.]

25. Show that the generators $\nu_1, \nu_2, \dots, \nu_n$ of $\ker\varphi$ have corresponding relations matrix

$$\begin{pmatrix} x - a_{11} & -a_{21} & \dots & -a_{n1} \\ -a_{12} & x - a_{22} & \dots & -a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{1n} & -a_{2n} & \dots & x - a_{nn} \end{pmatrix} = xI - A^t,$$

where A^t is the transpose of A. Conclude that Theorem 21 and the algorithm for determining the invariant factors of A follows by Exercises 16 to 19 in the previous section (note that the row and column operations necessary to diagonalize this relations matrix are the column and row operations necessary to diagonalize the matrix in Theorem 21, which explains why the invariant factor algorithm keeps track of the *row* operations used).

12.3 THE JORDAN CANONICAL FORM

We continue with the notation in the previous section: F is a field, $F[x]$ is the ring of polynomials in x with coefficients in F, V is a finite dimensional vector space over F of dimension n, T is a fixed linear transformation of V by which we make V into an $F[x]$-module, and A is an $n \times n$ matrix with coefficients in F. Recall that once a basis for V has been fixed any linear transformation T defines a matrix A and conversely any matrix A defines a linear transformation T.

In the previous section we used the invariant factor form of the Fundamental Theorem for finitely generated modules over the Principal Ideal Domain $F[x]$ to obtain the rational canonical form for such a linear transformation T and the rational canonical form for such an $n \times n$ matrix A. In this section we use the elementary divisor form of the Fundamental Theorem to obtain the *Jordan canonical form*. We shall see that matrices in this canonical form are as close to being diagonal matrices as possible, so the matrices are simpler than in the rational canonical form (but we lose some of the "rationality" results).

The elementary divisors of a module are the prime power divisors of its invariant factors (this was Corollary 10). For the $F[x]$-module V the invariant factors were monic polynomials $a_1(x), a_2(x), \dots, a_m(x)$ of degree at least one (with $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$), so the associated elementary divisors are the powers of the irreducible polynomial factors of these polynomials. These polynomials are only defined up to multiplication by a unit and, as in the case of the invariant factors, we can specify them uniquely by requiring that they be monic.

To obtain the simplest possible elementary divisors we shall assume that the polynomials $a_1(x), a_2(x), \dots, a_m(x)$ factor completely into linear factors, i.e., that the elementary divisors of V are powers $(x - \lambda)^k$ of linear polynomials. Since the product of the elementary divisors is the characteristic polynomial, this is equivalent to the assumption that the field F contains all the eigenvalues of the linear transformation T (equivalently, of the matrix A representing the linear transformation T).

Under this assumption on F, it follows immediately from Theorem 6 that V is the direct sum of finitely many cyclic $F[x]$-modules of the form $F[x]/(x - \lambda)^k$ where $\lambda \in F$ is one of the eigenvalues of T, corresponding to the elementary divisors of V.

We now choose a vector space basis for each of the direct summands corresponding to the elementary divisors of V for which the corresponding matrix for T is particularly simple. Recall that by definition of the $F[x]$-module structure the linear transformation T acting on V is the element x acting by multiplication on each of the direct summands $F[x]/(x - \lambda)^k$.

Consider the elements

$$(\bar{x} - \lambda)^{k-1},\ (\bar{x} - \lambda)^{k-2}, \dots,\ \bar{x} - \lambda,\ 1,$$

in the quotient $F[x]/(x-\lambda)^k$. Expanding each of these polynomials in $\bar{x}$ we see that the matrix relating these elements to the F-basis $\bar{x}^{k-1}, \bar{x}^{k-2}, \ldots, \bar{x}, 1$ of $F[x]/(x-\lambda)^k$ is upper triangular with 1's along the diagonal. Since this is an invertible matrix (having determinant 1), it follows that the elements above are an F-basis for $F[x]/(x-\lambda)^k$. With respect to this basis the linear transformation of multiplication by x acts in a particularly simple manner (note that $x = \lambda + (x - \lambda)$ and that $(\bar{x}-\lambda)^k = 0$ in the quotient):

$$x : \quad \begin{array}{lcl} (\bar{x}-\lambda)^{k-1} & \mapsto & \lambda\cdot(\bar{x}-\lambda)^{k-1} + (\bar{x}-\lambda)^{k} = \lambda\cdot(\bar{x}-\lambda)^{k-1} \\ (\bar{x}-\lambda)^{k-2} & \mapsto & \lambda\cdot(\bar{x}-\lambda)^{k-2} + (\bar{x}-\lambda)^{k-1} \\ & \vdots & \\ \bar{x}-\lambda & \mapsto & \lambda\cdot(\bar{x}-\lambda) + (\bar{x}-\lambda)^{2} \\ 1 & \mapsto & \lambda\cdot 1 + (\bar{x}-\lambda). \end{array}$$

With respect to this basis, the matrix for multiplication by x is therefore

$$\begin{pmatrix} \lambda & 1 & & & \\ & \lambda & \ddots & & \\ & & \ddots & 1 & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}$$

where the blank entries are all zero. Such matrices are given a name:

Definition. The $k \times k$ matrix with λ along the main diagonal and 1 along the first superdiagonal depicted above is called the *$k \times k$ elementary Jordan matrix with eigenvalue λ* or the *Jordan block of size k with eigenvalue λ*.

Applying this to each of the cyclic factors of V in its elementary divisor decomposition we obtain a vector space basis for V with respect to which the linear transformation T has as matrix the direct sum of the Jordan blocks corresponding to the elementary divisors of V, i.e., is block diagonal with Jordan blocks along the diagonal:

$$\begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_t \end{pmatrix}.$$

Notice that this matrix is uniquely determined up to permutation of the blocks along the diagonal by the elementary divisors of the $F[x]$-module V and conversely, by Theorem 9, the list of elementary divisors uniquely determines the module V up to $F[x]$-module isomorphism.

Definition.

(1) A matrix is said to be in *Jordan canonical form* if it is a block diagonal matrix with Jordan blocks along the diagonal.

(2) A *Jordan canonical form* for a linear transformation T is a matrix representing T which is in Jordan canonical form.

We have proved that any linear transformation T has a Jordan canonical form. As in the case of the rational canonical form, it follows from the uniqueness of the elementary divisors that the Jordan canonical form is unique up to a permutation of the Jordan blocks along the diagonal (hence is called *the* Jordan canonical form for T). We summarize this in the following theorem.

Theorem 22. *(Jordan Canonical Form for Linear Transformations)* Let V be a finite dimensional vector space over the field F and let T be a linear transformation of V. Assume F contains all the eigenvalues of T.

(1) There is a basis for V with respect to which the matrix for T is in Jordan canonical form, i.e., is a block diagonal matrix whose diagonal blocks are the Jordan blocks for the elementary divisors of V.

(2) The Jordan canonical form for T is unique up to a permutation of the Jordan blocks along the diagonal.

As for the rational canonical form, the following theorem gives the corresponding statement for $n \times n$ matrices over F.

Theorem 23. *(Jordan Canonical Form for Matrices)* Let A be an $n \times n$ matrix over the field F and assume F contains all the eigenvalues of A.

(1) The matrix A is similar to a matrix in Jordan canonical form, i.e., there is an invertible $n \times n$ matrix P over F such that $P^{-1}AP$ is a block diagonal matrix whose diagonal blocks are the Jordan blocks for the elementary divisors of A.

(2) The Jordan canonical form for A is unique up to a permutation of the Jordan blocks along the diagonal.

The Jordan canonical form differs from a diagonal matrix only by the possible presence of some 1's along the first superdiagonal (and then only if there are Jordan blocks of size greater than one), hence is close to being a diagonal matrix. The following result shows in particular that the Jordan canonical form for a matrix A is as close to being a diagonal matrix as possible.

Corollary 24.

(1) If a matrix A is similar to a diagonal matrix D, then D is the Jordan canonical form of A.

(2) Two diagonal matrices are similar if and only if their diagonal entries are the same up to a permutation.

Proof: The first assertion is immediate from the uniqueness of Jordan canonical forms because a diagonal matrix is itself in Jordan form (with Jordan blocks of size 1). The uniqueness of the Jordan canonical form gives (2).

The next corollary gives a criterion to determine when a matrix A can be diagonalized.

Corollary 25. If A is an $n \times n$ matrix with entries from F and F contains all the eigenvalues of A, then A is similar to a diagonal matrix over F if and only if the minimal polynomial of A has no repeated roots.

Proof: Suppose A is similar to a diagonal matrix. The minimal polynomial of a diagonal matrix has no repeated roots (its roots are precisely the distinct elements along the diagonal). Since similar matrices have the same minimal polynomial it follows that the minimal polynomial for A has no repeated roots.

Conversely, suppose the minimal polynomial for A has no repeated roots and let B be the Jordan canonical form of A. The matrix B is a block diagonal matrix with elementary Jordan matrices down the diagonal. By the exercises at the end of the preceding section the minimal polynomial for B is the least common multiple of the minimal polynomials of the Jordan blocks. It is easy to see directly that a Jordan block of size k with eigenvalue λ has minimal polynomial $(x - \lambda)^k$ (note that this is immediate from the fact that each elementary Jordan matrix gives the action on a *cyclic* $F[x]$-submodule whose annihilator is $(x-\lambda)^k$). Since A and B have the same minimal polynomial, the least common multiple of the $(x-\lambda)^k$ cannot have any repeated roots. It follows that k must be 1, i.e., that each Jordan block must be of size one and B is a diagonal matrix.

Changing From One Canonical Form to Another

We continue to assume that the field F contains all the eigenvalues of T (or A) so both the rational and Jordan canonical forms exist over F. The process of passing from one form to the other is exactly the same algorithm described in Section 5.2 for finite abelian groups (where the elementary divisors were determined from the list of invariant factors and vice versa).

In brief summary, recall that the elementary divisors are the prime power divisors of the invariant factors. They are obtained from the invariant factors by writing each invariant factor as a product of distinct linear factors to powers; the resulting set of powers of linear polynomials is the set of elementary divisors. For example, if the invariant factors of T are

$$(x-1)(x-3)^3, \quad (x-1)(x-2)(x-3)^3, \quad (x-1)(x-2)^2(x-3)^3$$

then the elementary divisors are

$$(x-1), \quad (x-3)^3, \quad (x-1), \quad (x-2), \quad (x-3)^3, \quad (x-1), \quad (x-2)^2, \quad (x-3)^3.$$

The largest invariant factor is the product of the largest of the distinct prime powers among the elementary divisors, the next largest invariant factor is the product of the largest of the distinct prime powers among the remaining elementary divisors, and so on. Given a list of elementary divisors we can find the list of invariant factors by first arranging the elementary divisors into n separate lists, one for each eigenvalue. In each of these n lists arrange the polynomials in increasing (i.e., nondecreasing) degree. Next arrange for all n lists to have the same length by appending an appropriate number of the constant polynomial 1. Now form the i^{th} invariant factor by taking the product of the i^{th} polynomial in each of these lists. For example, if the elementary divisors of T are

$$(x-1)^3,\ (x+4),\ (x+4)^2,\ (x-5)^2,\ (x-1)^5,\ (x-1)^3,\ (x-5)^3,\ (x-1)^4,\ (x+4)^3$$

then the intermediate lists are

$$\begin{array}{lllll} (1) & (x-1)^3, & (x-1)^3, & (x-1)^4, & (x-1)^5 \\ (2) & 1, & x+4, & (x+4)^2, & (x+4)^3 \\ (3) & 1, & 1, & (x-5)^2, & (x-5)^3 \end{array}$$

so the list of invariant factors is

$$(x-1)^3, \quad (x-1)^3(x+4), \quad (x-1)^4(x+4)^2(x-5)^2, \quad (x-1)^5(x+4)^3(x-5)^3.$$

Elementary Divisor Decomposition Algorithm: Converting to Jordan Canonical Forms

Theorem 21 indicates a computational procedure to determine the invariant factors of any given matrix A. Factorization of these invariant factors produces the elementary divisors of A, hence determines the Jordan canonical form for A as above.

The Invariant Factor Decomposition Algorithm following Theorem 21 starts with a basis $e_1, \dots, e_n$ for V and produces a set $f_1, \dots, f_m$ of elements of V which are $F[x]$-module generators for the cyclic factors in the invariant factor decomposition of V (with annihilators $(a_1(x)), \dots, (a_m(x))$, respectively). Since the elementary divisor decomposition is obtained from the invariant factor decomposition by applying the Chinese Remainder Theorem to the cyclic modules $F[x]/(a_i(x))$, this gives a set of $F[x]$-module generators for the cyclic factors in the elementary divisor decomposition of V. These elements then give rise to an explicit vector space basis for V with respect to which the linear transformation corresponding to A is in Jordan canonical form (equivalently, an explicit matrix P such that $P^{-1}AP$ is in Jordan canonical form). As for the Invariant Factor Decomposition Algorithm we state the result first in the general context of decomposing a vector space and then describe the algorithm to convert a given $n \times n$ matrix A to Jordan canonical form.

Explicit numerical examples of this algorithm are given later in Examples 2 and 3.

Elementary Divisor Decomposition Algorithm

(1) to **(3)**: The first three steps in the algorithm are those from the Invariant Factor Decomposition Algorithm following Theorem 21.

(4) For each invariant factor $a(x)$ computed for A write

$$a(x) = (x-\lambda_1)^{\alpha_1}(x-\lambda_2)^{\alpha_2}\dots(x-\lambda_s)^{\alpha_s}$$

where $\lambda_1, \dots, \lambda_s \in F$ are distinct. Let $f \in V$ be the $F[x]$-module generator for the cyclic factor corresponding to the invariant factor $a(x)$ computed in (3). Then the elements

$$\frac{a(x)}{(x-\lambda_1)^{\alpha_1}}f, \quad \frac{a(x)}{(x-\lambda_2)^{\alpha_2}}f, \quad \dots, \quad \frac{a(x)}{(x-\lambda_s)^{\alpha_s}}f$$

(note that the $\dfrac{a(x)}{(x-\lambda_i)^{\alpha_i}} \in F[x]$ are polynomials) are $F[x]$-module generators for the cyclic factors of V corresponding to the elementary divisors

$$(x-\lambda_1)^{\alpha_1}, \quad (x-\lambda_2)^{\alpha_2}, \quad \ldots, \quad (x-\lambda_s)^{\alpha_s},$$

respectively.

(5) If $g_i = \dfrac{a(x)}{(x-\lambda_i)^{\alpha_i}} f$ is the $F[x]$-module generator for the cyclic factor of V corresponding to the elementary divisor $(x-\lambda_i)^{\alpha_i}$ then the corresponding *vector space* basis for this cyclic factor of V is given by the elements

$$(T-\lambda_i)^{\alpha_i-1}g_i, \quad (T-\lambda_i)^{\alpha_i-2}g_i, \quad \ldots, \quad (T-\lambda_i)g_i, \quad g_i.$$

(6) Write the k^{th} element of the vector space basis computed in (5) in terms of the original vector space basis $[e_1, e_2, \ldots, e_n]$ for V and use the coordinates for the k^{th} column of an $n \times n$ matrix P. Then $P^{-1}AP$ is in Jordan canonical form (with Jordan blocks appearing in the order used in (5) for the cyclic factors of V).

Converting an $n \times n$ Matrix to Jordan Canonical Form

(1) to **(2)**: The first two steps are those from the algorithm for Converting an $n \times n$ matrix to Rational Canonical Form following Theorem 21.

(3) When $xI - A$ has been diagonalized to the form in Theorem 21 the first $n-m$ columns of the matrix P' are 0 (providing a useful numerical check on the computations) and the remaining m columns of P' are nonzero. For each successive $i = 1, 2, \ldots, m$:

(a) Factor the i^{th} nonconstant diagonal element (which is of degree d_i):

$$a(x) = (x-\lambda_1)^{\alpha_1}(x-\lambda_2)^{\alpha_2}\ldots(x-\lambda_s)^{\alpha_s}$$

where $\lambda_1, \ldots, \lambda_s \in F$ are distinct (here $a(x) = a_i(x)$ is the i^{th} nonconstant diagonal element and s depends on i).

(b) Multiply the i^{th} nonzero column of P' successively by the d_i matrices:

$$\begin{array}{lll}
(A-\lambda_1 I)^{\alpha_1-1}(A-\lambda_2 I)^{\alpha_2} & & \ldots(A-\lambda_s I)^{\alpha_s} \\
(A-\lambda_1 I)^{\alpha_1-2}(A-\lambda_2 I)^{\alpha_2} & & \ldots(A-\lambda_s I)^{\alpha_s} \\
 & \vdots & \\
(A-\lambda_1 I)^{0} & (A-\lambda_2 I)^{\alpha_2} & \ldots(A-\lambda_s I)^{\alpha_s} \\
 & & \\
(A-\lambda_1 I)^{\alpha_1} & (A-\lambda_2 I)^{\alpha_2-1} & \ldots(A-\lambda_s I)^{\alpha_s} \\
(A-\lambda_1 I)^{\alpha_1} & (A-\lambda_2 I)^{\alpha_2-2} & \ldots(A-\lambda_s I)^{\alpha_s} \\
 & \vdots & \\
(A-\lambda_1 I)^{\alpha_1} & (A-\lambda_2 I)^{0} & \ldots(A-\lambda_s I)^{\alpha_s} \\
 & \vdots &
\end{array}$$

$$\begin{gathered}\vdots\\ (A-\lambda_1 I)^{\alpha_1}(A-\lambda_2 I)^{\alpha_2}\dots(A-\lambda_s I)^{\alpha_s-1}\\ (A-\lambda_1 I)^{\alpha_1}(A-\lambda_2 I)^{\alpha_2}\dots(A-\lambda_s I)^{\alpha_s-2}\\ \vdots\\ (A-\lambda_1 I)^{\alpha_1}(A-\lambda_2 I)^{\alpha_2}\dots(A-\lambda_s I)^{0}.\end{gathered}$$

(c) Use the column vectors resulting from (b) (in that order) as the next d_i columns of an $n \times n$ matrix P.

Then $P^{-1}AP$ is in Jordan canonical form (whose Jordan blocks correspond to the ordering of the factors in (a)).

Examples

We can use Jordan canonical forms to carry out the same analysis of matrices that we did as examples of the use of rational canonical forms. In some instances, when the field is enlarged, the number of similarity classes increases (the number of similarity classes can never decrease when we extend the field by Corollary 18(2)).

(1) Let A, B and C be the matrices in Example 1 of the previous section and let $F = \mathbb{Q}$. Note that $\mathbb{Q}$ contains all the eigenvalues for these matrices. Since we have already determined the invariant factors of these matrices we can immediately obtain their elementary divisors. The elementary divisors of A are $x-2$, $x-2$ and $x-3$ and the elementary divisors of B and C are $(x-2)^2$ and $x-3$ so the respective Jordan canonical forms are:

$$\begin{pmatrix}2&0&0\\0&2&0\\0&0&3\end{pmatrix}\quad\begin{pmatrix}2&1&0\\0&2&0\\0&0&3\end{pmatrix}\quad\begin{pmatrix}2&1&0\\0&2&0\\0&0&3\end{pmatrix}.$$

Notice that A is similar to a diagonal matrix but, by Corollary 25, B and C are not.

(2) For the matrix A, we determined in Example 2 of the previous section that $f_1 = -7e_1 + 7e_2 + e_3$ and $f_2 = -e_1 + e_2$ were $\mathbb{Q}[x]$-module generators for the two cyclic factors of V in its invariant factor decomposition, corresponding to the invariant factors $x-2$ and $(x-2)(x-3)$, respectively. Using the first algorithm described above, the elements f_1, $(x-3)f_2$ and $(x-2)f_2$ are therefore $\mathbb{Q}[x]$-module generators for the three cyclic factors of V in its elementary divisor decomposition, corresponding to the elementary divisors $x-2$, $x-2$, and $x-3$. An easy computation shows that these are the elements $-7e_1 + 7e_2 + e_3$, $-e_1$ and $-2e_1 + e_2$, respectively. Then the matrix

$$P = \begin{pmatrix}-7&-1&-2\\7&0&1\\1&0&0\end{pmatrix}$$

conjugates A into its Jordan canonical form:

$$P^{-1}AP = \begin{pmatrix}2&0&0\\0&2&0\\0&0&3\end{pmatrix},$$

as one easily checks.

The columns of this matrix can also be obtained following the second algorithm above, using the nonzero columns of the matrix P' computed in Example 2 of the previous section:

$$(A-2I)^0\begin{pmatrix}-7\\7\\1\end{pmatrix}=\begin{pmatrix}-7\\7\\1\end{pmatrix}$$

and

$$(A-2I)^0(A-3I)^1\begin{pmatrix}-1\\1\\0\end{pmatrix}=\begin{pmatrix}-1\\0\\0\end{pmatrix},\quad (A-2I)^1(A-3I)^0\begin{pmatrix}-1\\1\\0\end{pmatrix}=\begin{pmatrix}-2\\1\\0\end{pmatrix},$$

respectively, which again gives the matrix P.

(3) For the 4×4 matrix D of Example 3 of the previous section, the invariant factors were $(x-1)^2$, $(x-1)^2$, with corresponding $\mathbb{Q}[x]$-module generators $f_1 = e_1$ and $f_2 = e_2$, respectively. These are also the elementary divisors for this matrix. The corresponding vector space bases for these two factors are given by $(T-1)f_1, f_1$ and $(T-1)f_2, f_2$, respectively. An easy computation shows these are the elements $2e_2 + e_3, e_1$ and $2e_1 - e_2 + e_4, e_2$, respectively. Then the matrix

$$P=\begin{pmatrix}0&1&2&0\\2&0&-2&1\\1&0&0&0\\0&0&1&0\end{pmatrix}$$

conjugates D into its Jordan canonical form:

$$P^{-1}DP=\begin{pmatrix}1&1&0&0\\0&1&0&0\\0&0&1&1\\0&0&0&1\end{pmatrix}$$

as can easily be checked.

The columns of this matrix can also be obtained following the second algorithm above, using the nonzero columns of the matrix P' computed in Example 3 of the previous section:

$$(D-I)^1\begin{pmatrix}1\\0\\0\\0\end{pmatrix}=\begin{pmatrix}0\\2\\1\\0\end{pmatrix},\quad (D-I)^0\begin{pmatrix}1\\0\\0\\0\end{pmatrix}=\begin{pmatrix}1\\0\\0\\0\end{pmatrix},$$

and

$$(D-I)^1\begin{pmatrix}0\\1\\0\\0\end{pmatrix}=\begin{pmatrix}2\\-2\\0\\1\end{pmatrix},\quad (D-I)^0\begin{pmatrix}0\\1\\0\\0\end{pmatrix}=\begin{pmatrix}0\\1\\0\\0\end{pmatrix},$$

respectively, which again gives the matrix P.

(4) The set of similarity classes of 6×6 matrices with entries from $\mathbb{C}$ with characteristic polynomial $(x^4-1)(x^2-1)$ consists of the 4 classes represented by the rational canonical forms in the preceding set of examples (there are no additional lists of invariant factors over $\mathbb{C}$). Their Jordan canonical forms cannot all be written over $\mathbb{Q}$, however. For instance, if the invariant factors are

$$(x-1)(x+1) \quad\text{and}\quad (x-1)(x+1)(x^2+1)$$

then the elementary divisors are

$$x-1, \quad x+1, \quad x-1, \quad x+1, \quad x-i, \quad x+i,$$

where i is a square root of -1 in $\mathbb{C}$, so the Jordan form for this matrix is a diagonal matrix with diagonal entries $1, 1, -1, -1, i, -i$.

(5) In contrast, the set of similarity classes of 3×3 matrices, A, over $\mathbb{C}$ satisfying $A^6 = I$ is considerably larger than that over $\mathbb{Q}$. If A is any such matrix, $m_A(x) \mid x^6 - 1$ so since the latter polynomial has no repeated roots in $\mathbb{C}$, the minimal polynomial of A has no repeated roots. By Corollary 25 the Jordan canonical form of A is a diagonal matrix. Since this diagonal matrix has the same minimal polynomial, its 6th power is also the identity, and so each diagonal entry is a 6th root of unity. For each list $\zeta_1, \zeta_2, \zeta_3$ of 6th roots of unity we obtain a Jordan canonical form, and two such forms are the same (i.e., give rise to similar matrices) if and only if the lists are permuted versions of each other. One finds that there are, up to similarity, 56 classes of such A's.

EXERCISES

1. Suppose the vector space V is the direct sum of cyclic $F[x]$-modules whose annihilators are $(x+1)^2$, $(x-1)(x^2+1)^2$, (x^4-1) and $(x+1)(x^2-1)$. Determine the invariant factors and elementary divisors for V.

2. Prove that if $\lambda_1, \dots, \lambda_n$ are the eigenvalues of the $n \times n$ matrix A then $\lambda_1^k, \dots, \lambda_n^k$ are the eigenvalues of A^k for any $k \geq 0$.

3. Use the method of Example 2 above to determine explicit matrices P_1 and P_2 with $P_1^{-1} B P_1$ and $P_2^{-1} C P_2$ in Jordan canonical form. Use this to explicitly construct a matrix Q which conjugates B into C (proving directly that these matrices are similar).

4. Prove that the Jordan canonical form for the matrix

$$\begin{pmatrix} 9 & 4 & 5 \\ -4 & 0 & -3 \\ -6 & -4 & -2 \end{pmatrix}$$

is that stated at the beginning of this chapter. Explicitly determine a matrix P which conjugates this matrix to its Jordan canonical form. Explain why this matrix cannot be diagonalized.

5. Compute the Jordan canonical form for the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -2 \\ 0 & 1 & 3 \end{pmatrix}.$$

6. Determine which of the following matrices are similar:

$$\begin{pmatrix} -1 & 4 & -4 \\ 2 & -1 & 3 \\ 0 & -4 & 3 \end{pmatrix} \quad \begin{pmatrix} -3 & -4 & 0 \\ 2 & 3 & 0 \\ 8 & 8 & 1 \end{pmatrix} \quad \begin{pmatrix} -3 & 2 & -4 \\ 2 & 1 & 0 \\ 3 & -1 & 3 \end{pmatrix} \quad \begin{pmatrix} -1 & 4 & -4 \\ 0 & -3 & 2 \\ 0 & -4 & 3 \end{pmatrix}.$$

7. Determine the Jordan canonical forms for the following matrices:

$$\begin{pmatrix} 5 & 4 & 1 \\ -1 & 0 & 0 \\ -3 & -4 & 1 \end{pmatrix} \quad \begin{pmatrix} 3 & 4 & 2 \\ -2 & -3 & -1 \\ -4 & -4 & -3 \end{pmatrix}.$$

8. Prove that the matrices

$$A = \begin{pmatrix} 5 & 6 & 0 \\ -3 & -4 & 0 \\ -2 & 0 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 3 & -1 & 2 \\ -10 & 6 & -14 \\ -6 & 3 & -7 \end{pmatrix}$$

are similar. Prove that both A and B can be diagonalized and determine explicit matrices P_1 and P_2 with $P_1^{-1}AP_1$ and $P_2^{-1}BP_2$ in diagonal form.

9. Prove that the matrices

$$A = \begin{pmatrix} -8 & -10 & -1 \\ 7 & 9 & 1 \\ 3 & 2 & 0 \end{pmatrix} \quad B = \begin{pmatrix} -3 & 2 & -4 \\ 4 & -1 & 4 \\ 4 & -2 & 5 \end{pmatrix}$$

both have $(x-1)^2(x+1)$ as characteristic polynomial but that one can be diagonalized and the other cannot. Determine the Jordan canonical form for both matrices.

10. Find all Jordan canonical forms of 2×2, 3×3 and 4×4 matrices over $\mathbb{C}$.

11. Verify that the characteristic polynomial of

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & -2 & 0 & 1 \\ -2 & 0 & -1 & -2 \end{pmatrix}$$

is a product of linear factors over $\mathbb{Q}$. Determine the rational and Jordan canonical forms for A over $\mathbb{Q}$.

12. Determine the Jordan canonical form for the matrix

$$\begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

13. Determine the Jordan canonical form for the matrix

$$\begin{pmatrix} 3 & 0 & -2 & -3 \\ 4 & -8 & 14 & -15 \\ 2 & -4 & 7 & -7 \\ 0 & 2 & -4 & 3 \end{pmatrix}.$$

14. Prove that the matrices

$$A = \begin{pmatrix} 2 & 0 & 0 & 0 \\ -4 & -1 & -4 & 0 \\ 2 & 1 & 3 & 0 \\ -2 & 4 & 9 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 5 & 0 & -4 & -7 \\ 3 & -8 & 15 & -13 \\ 2 & -4 & 7 & -7 \\ 1 & 2 & -5 & 1 \end{pmatrix}$$

are similar.

15. Prove that the matrices

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 5 & 2 & -8 & -8 \\ -6 & -3 & 8 & 8 \\ -3 & -1 & 3 & 4 \\ 3 & 1 & -4 & -5 \end{pmatrix}$$

both have characteristic polynomial $(x-3)(x+1)^3$. Determine whether they are similar and determine the Jordan canonical form for each matrix.

16. Determine the Jordan canonical form for the matrix

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

and determine a matrix P which conjugates this matrix into its Jordan canonical form.

17. Prove that any matrix A is similar to its transpose A^t.

18. Determine all possible Jordan canonical forms for a linear transformation with characteristic polynomial $(x-2)^3(x-3)^2$.

19. Prove that all $n \times n$ matrices with characteristic polynomial $f(x)$ are similar if and only if $f(x)$ has no repeated factors in its unique factorization in $F[x]$.

20. Show that the following matrices are similar in $M_p(\mathbb{F}_p)$ ($p \times p$ matrices with entries from $\mathbb{F}_p$):

$$\begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 1 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

21. Show that if $A^2 = A$ then A is similar to a diagonal matrix which has only 0's and 1's along the diagonal.

22. Prove that an $n \times n$ matrix A with entries from $\mathbb{C}$ satisfying $A^3 = A$ can be diagonalized. Is the same statement true over *any* field F?

23. Suppose A is a 2×2 matrix with entries from $\mathbb{Q}$ for which $A^3 = I$ but $A \neq I$. Write A in rational canonical form and in Jordan canonical form viewed as a matrix over $\mathbb{C}$.

24. Prove there are no 3×3 matrices A over $\mathbb{Q}$ with $A^8 = I$ but $A^4 \neq I$.

25. Determine the Jordan canonical form for the $n \times n$ matrix over $\mathbb{Q}$ whose entries are all equal to 1.

26. Determine the Jordan canonical form for the $n \times n$ matrix over $\mathbb{F}_p$ whose entries are all equal to 1 (the answer depends on whether or not p divides n).

27. Determine the Jordan canonical form for the $n \times n$ matrix over $\mathbb{Q}$ whose entries are all equal to 1 except that the entries along the main diagonal are all equal to 0.

28. Determine the Jordan canonical form for the $n \times n$ matrix over $\mathbb{F}_p$ whose entries are all equal to 1 except that the entries along the main diagonal are all equal to 0.

The direct sum of the cyclic submodules of V corresponding to all the elementary divisors of V which are powers of the same $x - \lambda$ is called the *generalized eigenspace of T* corresponding to the eigenvalue λ. Note that this is the p-primary component of V for the prime $p = x - \lambda$ of $F[x]$ and consists of the elements of V which are annihilated by some power of the linear transformation $T - \lambda$. The matrix for T on the generalized eigenspace for λ is the block diagonal matrix of all Jordan blocks for T with the same eigenvalue λ.

29. Suppose V_i is the generalized eigenspace of T corresponding to eigenvalue λ_i. For any $k \geq 0$, prove that the nullity of $T - \lambda_i$ on the subspace $(T - \lambda_i)^k V_i$ is the same as the nullity of $T - \lambda_i$ on $(T - \lambda_i)^k V$ and equals the number of Jordan blocks of T having eigenvalue λ_i and size greater than k (so for $k = 0$ this gives the number of Jordan blocks).

30. Let λ be an eigenvalue of the linear transformation T on the finite dimensional vector space V over the field F. Let $r_k = \dim_F (T - \lambda)^k V$ be the rank of the linear transformation $(T - \lambda)^k$ on V. For any $k \geq 1$, prove that $r_{k-1} - 2r_k + r_{k+1}$ is the number of Jordan blocks of T corresponding to λ of size k [use Exercise 12 in Section 1]. (This gives an efficient method for determining the Jordan canonical form for T by computing the ranks of the matrices $(A - \lambda I)^k$ for a matrix A representing T, cf. Exercise 31(a) in Section 11.2.)

31. Let N be an $n \times n$ matrix with coefficients in the field F. The matrix N is said to be *nilpotent* if some power of N is the zero matrix, i.e., $N^k = 0$ for some k. Prove that any nilpotent matrix is similar to a block diagonal matrix whose blocks are matrices with 1's along the first superdiagonal and 0's elsewhere.

32. Prove that if N is an $n \times n$ nilpotent matrix then in fact $N^n = 0$.

33. Let A be a strictly upper triangular $n \times n$ matrix (all entries on and below the main diagonal are zero). Prove that A is nilpotent.

34. Prove that the trace of a nilpotent $n \times n$ matrix is 0 (recall the trace of a matrix is the sum of the diagonal elements).

35. For $0 \leq i \leq n$, let d_i be the g.c.d. of the determinants of all the $i \times i$ minors of $xI - A$, for A as in Theorem 21 (take the 0×0 minor to be 1). Prove that the i^{th} element along the diagonal of the Smith Normal Form for A is d_i / d_{i-1}. This gives the invariant factors for A. [Show these g.c.d.s do not change under elementary row and column operations.]

36. Let $V = \mathbb{C}^n$ be the usual n-dimensional vector space of n-tuples $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ of complex numbers. Let T be the linear transformation defined by setting $T(\alpha_1, \alpha_2, \ldots, \alpha_n)$ equal to $(0, \alpha_1, \alpha_2, \ldots, \alpha_{n-1})$. Determine the Jordan canonical form for T.

37. Let J be a Jordan block of size n with eigenvalue λ over $\mathbb{C}$.

(a) Prove that the Jordan canonical form for the matrix J^2 is the Jordan block of size n with eigenvalue λ^2 if $\lambda \neq 0$.

(b) If $\lambda = 0$ prove that the Jordan canonical form for J^2 has two blocks (with eigenvalues 0) of size $\frac{n}{2}, \frac{n}{2}$ if n is even and of size $\frac{n-1}{2}, \frac{n+1}{2}$ if n is odd.

38. Determine necessary and sufficient conditions for a matrix $A \in M_n(\mathbb{C})$ to have a square root, i.e., for there to exist another matrix $B \in M_n(\mathbb{C})$ such that $A = B^2$. [Suppose B is in Jordan canonical form and consider the Jordan canonical form for B^2 using the previous exercise.]

39. Let J be a Jordan block of size n with eigenvalue λ over a field F of characteristic 2. Determine the Jordan canonical form for the matrix J^2. Determine necessary and sufficient conditions for a matrix $A \in M_n(F)$ to have a square root, i.e., for there to exist another matrix $B \in M_n(F)$ such that $A = B^2$.

The remaining exercises explore functions (power series) of a matrix and introduce some applications of the Jordan canonical form to the theory of differential equations.

Throughout these exercises the matrices are assumed to be $n \times n$ matrices with entries from the field K, where K is either the real or complex numbers. Let

$$G(x) = \sum_{k=0}^{\infty} \alpha_k x^k$$

be a power series with coefficients from K. Let $G_N(x) = \sum_{k=0}^{N} \alpha_k x^k$ be the N^{th} partial sum of $G(x)$ and for each $A \in M_n(K)$ let $G_N(A)$ be the element of $M_n(K)$ obtained (as usual) by substituting A in this polynomial. For each fixed i, j we obtain a sequence of real or complex numbers c_{ij}^N, $N = 0, 1, 2, \dots$ by taking c_{ij}^N to be the i, j entry of the matrix $G_N(A)$. The series

$$G(A) = \sum_{k=0}^{\infty} \alpha_k A^k$$

is said to *converge* to the matrix C in $M_n(K)$ if for each $i, j \in \{1, 2, \dots, n\}$ the sequence c_{ij}^N, $N = 0, 1, 2, \dots$ converges to the i, j entry of C (in which case we write $G(A) = C$). Say $G(A)$ *converges* if there is some $C \in M_n(K)$ such that $G(A) = C$. If A is a 1×1 matrix, this is the usual notion of convergence of a series in K.

For $A = (a_{ij}) \in M_n(K)$ define

$$|| A || = \sum_{i,j=1}^{n} |a_{ij}|$$

i.e., $|| A ||$ is the sum of the absolute values of all the entries of A.

40. Prove that for all $A, B \in M_n(K)$ and all $\alpha \in K$
- **(a)** $|| A + B || \leq || A || + || B ||$
- **(b)** $|| AB || \leq || A || \cdot || B ||$
- **(c)** $|| \alpha A || = |\alpha| \cdot || A ||$.

41. Let R be the radius of convergence of the real or complex power series $G(x)$ (where $R = \infty$ if $G(x)$ converges for all $x \in K$).
- **(a)** Prove that if $|| A || < R$ then $G(A)$ converges.
- **(b)** Deduce that for *all* matrices A the following power series converge:

$$\sin(A) = A - \frac{A^3}{3!} + \frac{A^5}{5!} + \cdots + (-1)^k \frac{A^{2k+1}}{(2k+1)!} + \cdots$$

$$\cos(A) = I - \frac{A^2}{2!} + \frac{A^4}{4!} + \cdots + (-1)^k \frac{A^{2k}}{(2k)!} + \cdots$$

$$\exp(A) = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots + \frac{A^k}{k!} + \cdots$$

where I is the $n \times n$ identity matrix.

In view of applications to the theory of differential equations we introduce a variable t at this point, so that for $A \in M_n(K)$ the matrix At is obtained from A by multiplying each entry by t (which is the same as multiplying A by the "scalar" matrix tI). We obtain a function from a subset of K into $M_n(K)$ defined by $t \mapsto G(At)$ at all points t where the series $G(At)$ converges. In particular, $\sin(At)$, $\cos(At)$ and $\exp(At)$ converge for all $t \in K$.

42. Let P be a nonsingular $n \times n$ matrix.
- **(a)** Prove that $PG(At)P^{-1} = G(PAtP^{-1}) = G(PAP^{-1}t)$. (This implies that, up to a change of basis, it suffices to compute $G(At)$ for matrices A in canonical form). [Take limits of partial sums to get the first equality. The second equality is immediate because the matrix tI commutes with every matrix.]
- **(b)** Prove that if A is the direct sum of matrices $A_1, A_2, \dots, A_m$, then $G(At)$ is the direct sum of the matrices $G(A_1t), G(A_2t), \dots, G(A_mt)$.
- **(c)** Show that if Z is the diagonal matrix with entries $z_1, z_2, \dots, z_n$ then $G(Zt)$ is the diagonal matrix with entries $G(z_1t), G(z_2t), \dots, G(z_nt)$.

The matrix $\exp(A)$ defined in Exercise 41(b) is called the *exponential* of A and is often denoted by e^A. The next three exercises lead to a formula for the matrix $\exp(Jt)$, where J is an elementary Jordan matrix.

43. Prove that if A and B are *commuting* matrices then $\exp(A + B) = \exp(A)\exp(B)$. [Treat A and B as commuting indeterminates and deduce this by comparing the power series on the left hand side with the product of the two power series on the right hand side.]

44. Use the preceding exercise to show that if M is any matrix and λ is any element of K then

$$\exp(\lambda I t + M) = e^{\lambda t}\exp(M).$$

45. Let N be the $r \times r$ matrix with 1's on the first superdiagonal and zeros elsewhere. Compute the exponential of the following nilpotent $r \times r$ matrix:

$$\text{if } Nt = \begin{pmatrix} 0 & t & & & \\ & 0 & t & & \\ & & & \ddots & \\ & & & & t \\ & & & & 0 \end{pmatrix} \quad \text{then } \exp(Nt) = \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \cdots & \frac{t^{r-1}}{(r-1)!} \\ & 1 & t & \frac{t^2}{2!} & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & t & \frac{t^2}{2!} \\ & & & & 1 & t \\ & & & & & 1 \end{pmatrix}.$$

Deduce that if J is the $r \times r$ elementary Jordan matrix with eigenvalue λ then

$$\exp(Jt) = \begin{pmatrix} e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} & \cdots & \cdots & \frac{t^{r-1}}{(r-1)!}e^{\lambda t} \\ & e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} \\ & & & & e^{\lambda t} & te^{\lambda t} \\ & & & & & e^{\lambda t} \end{pmatrix}.$$

[To do the first part use the observation that since Nt is a nilpotent matrix, $\exp(Nt)$ is a *polynomial* in Nt, i.e., all but a finite number of the terms in the power series are zero. To compute the exponential of Jt write Jt as $\lambda It + Nt$ and use Exercise 44 with $M = Nt$.]

Let $A \in M_n(K)$ and let P be a change of basis matrix such that $P^{-1}AP$ is in Jordan canonical form. Suppose $P^{-1}AP$ is the sum of elementary Jordan matrices $J_1, \dots, J_m$. The preceding exercises (with $t = 1$) show that $\exp(A)$ can easily be found by writing $E = \exp(P^{-1}AP)$ as the direct sum of the matrices $\exp(J_1), \dots, \exp(J_m)$ and then changing the basis back again to obtain $\exp(A) = PEP^{-1}$.

46. For the 4×4 matrices D and P given in Example 3 of this section:

$$D = \begin{pmatrix} 1 & 2 & -4 & 4 \\ 2 & -1 & 4 & -8 \\ 1 & 0 & 1 & -2 \\ 0 & 1 & -2 & 3 \end{pmatrix} \qquad P = \begin{pmatrix} 0 & 1 & 2 & 0 \\ 2 & 0 & -2 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

show that

$$E = \begin{pmatrix} e & e & 0 & 0 \\ 0 & e & 0 & 0 \\ 0 & 0 & e & e \\ 0 & 0 & 0 & e \end{pmatrix} \quad \text{and} \quad \exp(D) = \begin{pmatrix} e & 2e & -4e & 4e \\ 2e & -e & 4e & -8e \\ e & 0 & e & -2e \\ 0 & e & -2e & 3e \end{pmatrix}.$$

47. Compute the exponential of each of the following matrices:

(a) the matrix A in Example 2 of this section

(b) the matrix in Exercise 4 (where you computed the Jordan canonical form and a change of basis matrix)

(c) the matrix in Exercise 16.

48. Show that $\exp(0) = I$ (here 0 is the zero matrix and I is the identity matrix). Deduce that $\exp(A)$ is nonsingular with inverse $\exp(-A)$ for all matrices $A \in M_n(K)$.

49. Prove that $\det(\exp(A)) = e^{\mathrm{tr}(A)}$, where $\mathrm{tr}(A)$ is the trace of A (the sum of the diagonal entries of A).

50. Fix any $A \in M_n(K)$. Prove that the map

$$K \to GL_n(K) \qquad \text{defined by} \qquad t \mapsto \exp(At)$$

is a group homomorphism (here K is the additive group of the field). (Note how this generalizes the familiar exponential map from K to $K^\times$, which is the $n = 1$ case. The subgroup $\{\exp(At) \mid t \in K\}$ is called a *1-parameter subgroup* of $GL_n(K)$. These subgroups and the exponential map play an important role in the theory of *Lie groups* — $GL_n(K)$ being a particular example of a Lie group.).

Let $G(x)$ be a power series having an infinite radius of convergence and fix a matrix $A \in M_n(K)$. The entries of the matrix $G(At)$ are K-valued functions of the variable t that are defined for all t. Let $c_{ij}(t)$ be the function of t in the i, j entry of $G(At)$. The *derivative* of $G(At)$ with respect to t, denoted by $\frac{d}{dt}G(At)$, is the matrix whose i, j entry is $\frac{d}{dt}c_{ij}(t)$ obtained by differentiating each of the entries of $G(At)$. In other words, if we identify $M_n(K)$ with K^{n^2} by considering each $n \times n$ matrix as an n^2-tuple, then $t \mapsto G(At)$ is a map from K to K^{n^2} (i.e., is a vector valued function of t) whose derivative is just the usual (componentwise) derivative of this vector valued function.

51. Establish the following properties of derivatives:

(a) If $G(x) = \sum_{k=0}^{\infty} \alpha_k x^k$ then $\frac{d}{dt}G(At) = A \sum_{k=1}^{\infty} k\alpha_k (At)^{k-1}$.

(b) If v is an $n \times 1$ matrix with (constant) entries from K then

$$\frac{d}{dt}(G(At)v) = \left(\frac{d}{dt}G(At)\right) v.$$

52. Deduce from part (a) of the preceding exercise that

$$\frac{d}{dt}\exp(At) = A\exp(At).$$

Now let $y_1(t), \ldots, y_n(t)$ be differentiable functions of the real variable t that are related by the following linear system of first order differential equations with constant coefficients $a_{ij} \in K$:

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n \\ y_2' &= a_{21}y_1 + a_{22}y_2 + \ldots + a_{2n}y_n \\ &\ \ \vdots \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \ldots + a_{nn}y_n \end{aligned} \tag{$*$}$$

(here the primes denote derivatives with respect to t). Let A be the matrix whose i, j entry is a_{ij}, so that $(*)$ may be written as

$$\begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix} = A \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

or, more succinctly, as $y' = Ay$, where y is the column vector of functions $y_1(t), \ldots, y_n(t)$.

An $n \times n$ matrix whose entries are functions of t and whose columns are independent solutions to the system $(*)$ is called a *fundamental matrix* of $(*)$. By the theory of differential equations, the set of vectors y that are solutions to the system $(*)$ form an n-dimensional vector space over K and so the columns of a fundamental matrix are a *basis for the vector space of all solutions to* $(*)$.

53. Prove that $\exp(At)$ is a fundamental matrix of $(*)$. Show also that if C is the $n \times 1$ constant vector whose entries are $y_1(0), \ldots, y_n(0)$ then $y(t) = \exp(At)C$ is the particular solution to the system $(*)$ satisfying the initial condition $y(0) = C$. (Note how this generalizes the 1-dimensional result that the single differential equation $y' = ay$ has e^{at} as a basis for the 1-dimensional space of solutions and the unique solution to this differential equation satisfying the initial condition $y(0) = c$ is $y = ce^{at}$.) [Use the preceding exercises.]

54. Prove that if M is a fundamental matrix of $(*)$ and if Q is a nonsingular matrix in $M_n(K)$, then MQ is also a fundamental matrix of $(*)$. [The columns of MQ are linear combinations of the columns of M.]

Now apply the preceding two exercises to solve some specific systems of differential equations as follows: given the matrix A in a system $(*)$, calculate a change of basis matrix P such that $B = P^{-1}AP$ is in Jordan canonical form. Then $\exp(At) = P\exp(Bt)P^{-1}$ is a fundamental matrix for $(*)$. By the preceding exercise, $P\exp(Bt)$ is also a fundamental matrix for $(*)$ and $\exp(Bt)$ can be calculated by the method described in the discussion following Exercise 45 (in particular, one does not have to find the inverse of the matrix P to obtain a fundamental matrix for $(*)$). Thus, for example, if $A = D$ and P are the matrices given in Exercise 46, then we saw that the Jordan canonical form for A is the matrix $B = P^{-1}AP$ consisting of two 2×2 Jordan blocks with eigenvalues 1. A fundamental matrix for the system $y' = Ay$ is therefore

$$P\exp(B) = \begin{pmatrix} 0 & 1 & 2 & 0 \\ 2 & 0 & -2 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} e^t & te^t & 0 & 0 \\ 0 & e^t & 0 & 0 \\ 0 & 0 & e^t & te^t \\ 0 & 0 & 0 & e^t \end{pmatrix} = \begin{pmatrix} 0 & e^t & 2e^t & 2te^t \\ 2e^t & 2te^t & -2e^t & e^t(1-2t) \\ e^t & te^t & 0 & 0 \\ 0 & 0 & e^t & te^t \end{pmatrix}.$$

Writing this out more explicitly, this shows that the general solution to the system of differential equations

$$\begin{aligned} y_1' &= y_1 + 2y_2 - 4y_3 + 4y_4 \\ y_2' &= 2y_1 - y_2 + 4y_3 - 8y_4 \\ y_3' &= y_1 + y_3 - 2y_4 \\ y_4' &= y_2 - 2y_3 + 3y_4 \end{aligned}$$

is given by

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \alpha_1 \begin{pmatrix} 0 \\ 2e^t \\ e^t \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} e^t \\ 2te^t \\ te^t \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 2e^t \\ -2e^t \\ 0 \\ e^t \end{pmatrix} + \alpha_4 \begin{pmatrix} 2te^t \\ e^t(1-2t) \\ 0 \\ te^t \end{pmatrix}$$

where $\alpha_1, \ldots, \alpha_4$ are arbitrary elements of the field K (this describes the 4-dimensional vector space of solutions).

55. In each of Parts (a) to (c) find a fundamental matrix for the system $(*)$, where the coefficient matrix A of $(*)$ is specified.

(a) A is the matrix in Part (a) of Exercise 47.

(b) A is the matrix in Part (b) of Exercise 47.

(c) A is the matrix in Part (c) of Exercise 47.

56. Consider the system $(*)$ whose coefficient matrix A is the matrix D listed in Exercise 46 and whose fundamental matrix was computed just before the preceding exercise. Find the particular solution to $(*)$ that satisfies the initial condition $y_i(0) = 1$ for $i = 1, 2, 3, 4$.

Next we explore a special case of $(*)$. Given the linear n^{th} *order* differential equation with constant coefficients

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = 0 \qquad (**)$$

(where $y^{(k)}$ is the k^{th} derivative of y and $y^{(0)} = y$) one can form a *system* of linear *first order* differential equations by letting $y_i = y^{(i-1)}$ for $1 \leq i \leq n$ (the coefficient matrix of this system is described in the next exercise). A basis for the n-dimensional vector space of solutions to the n^{th} order equation $(**)$ may then obtained from a fundamental matrix for the linear system. Specifically, in each of the $n \times 1$ columns of functions in a fundamental matrix for the system, the $1, 1$ entry is a solution to $(**)$ and so the n functions in the first row of the fundamental matrix for the system form a basis for the solutions to $(**)$.

57. Prove that the matrix, A, of coefficients of the system of n first order equations obtained from $(**)$ is the transpose of the companion matrix of the polynomial $x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$.

58. Use the above methods to find a basis for the vector space of solutions to the following differential equations

(a) $y''' - 3y' + 2y = 0$

(b) $y'''' + 4y''' + 6y'' + 4y' + y = 0$.

A system of differential equations

$$\begin{aligned} y_1' &= F_1(y_1, y_2, \ldots, y_n) \\ y_2' &= F_2(y_1, y_2, \ldots, y_n) \\ &\ \ \vdots \\ y_n' &= F_n(y_1, y_2, \ldots, y_n) \end{aligned}$$

where $F_1, F_2, \ldots, F_n$ are functions of n variables, is called an *autonomous* system and it will be written more succinctly as $y' = F(y)$, where $F = (F_1, \ldots, F_n)$. (The expression autonomous means "independent of time" and it indicates that the variable t — which may be thought of as a time variable — does not appear explicitly on the right hand side.) The system $(*)$ is the special type of autonomous system in which each F_i is a linear function. In many instances it is desirable to analyze the behavior of solutions to an autonomous system of differential equations without explicitly finding these solutions (indeed, it is unlikely that it will be possible to find explicit solutions for a given nonlinear system). This investigation falls under the rubric "qualitative analysis" of autonomous differential equations and the rudiments of this study are often treated in basic calculus courses for 1×1 systems. The first step in a qualitative analysis of an $n \times n$ autonomous system is to find the *steady states*, namely the constant solutions (these are called steady states since they do not change with t). Note that a constant function $y = c$, where c is the $n \times 1$ constant vector with entries $c_1, \dots, c_n$, is a solution to $y' = F(y)$ if and only if

$$c_i' = 0 = F_i(c_1, \dots, c_n) \quad \text{for } i = 1, 2, \dots, n,$$

so the steady states are found by computing the zeros of F (in the case of a nonlinear system this may require numerical methods). Next, given the initial value of some solution, one wishes to analyze the behavior of this solution as $t \to \infty$. This is called the *asymptotic behavior* of the solution. Again, it may not be possible to find the solution explicitly, although by the general theory of differential equations a solution to the initial value problem is unique provided the functions F_i are differentiable. A steady state $y = c$ is called *globally asymptotically stable* if every solution tends to c as $t \to \infty$, i.e., for any solution $y(t)$ we have $\lim_{t\to\infty} y_i(t) = c_i$ for all $i = 1, 2, \dots, n$.

In the case of the linear autonomous system $(*)$ the solutions form a vector space, so the only constant solution is the zero solution. The next exercise gives a *sufficient* condition for zero to be globally asymptotically stable and it gives one example of how the behavior of a linear system may be analyzed in terms of the eigenvalues of its coefficient matrix. Nonlinear systems can be approximated by linear systems in some neighborhood of a steady state by considering $y' = Ty$, where $T = \left(\dfrac{\partial F_i}{\partial y_j}\right)$ is the $n \times n$ Jacobian matrix of F evaluated at the steady state point. In this way the analysis of linear systems plays an important role in the local analysis of general autonomous systems.

59. Prove that the solution of $(*)$ given by $y_i(t) = 0$ for all $i \in \{1, \dots, n\}$ (i.e., the zero solution) is globally asymptotically stable if all the eigenvalues of A have negative real parts. [For those unfamiliar with the behavior of the complex exponential function, assume all eigenvalues are real (hence are negative real numbers). Use the explicit nature of the solutions to show that they all tend to zero as $t \to \infty$.]

Part IV

FIELD THEORY AND GALOIS THEORY

The previous sections have developed the theory of some of the basic algebraic structures of groups, rings and fields. The next two chapters consider properties of fields, particularly fields which arise from trying to solve equations (such as the simple equation $x^2 + 1 = 0$), and fields which naturally arise in trying to perform "arithmetic" (adding, subtracting, multiplying and dividing). The elegant and beautiful Galois Theory relates the structure of *fields* to certain related *groups* and is one of the basic algebraic tools. Applications include solutions of classical compass and straightedge construction questions, finite fields and Abel's famous theorem on the insolvability (by radicals) of the general quintic polynomial.

CHAPTER 13

Field Theory

13.1 BASIC THEORY OF FIELD EXTENSIONS

Recall that a field F is a commutative ring with identity in which every nonzero element has an inverse. Equivalently, the set $F^\times = F - \{0\}$ of nonzero elements of F is an abelian group under multiplication.

One of the first invariants associated with any field F is its *characteristic*, defined as follows: If 1_F denotes the identity of F, then F contains the elements 1_F, $1_F + 1_F$, $1_F + 1_F + 1_F, \dots$ of the additive subgroup of F generated by 1_F, which may not all be distinct. For n a positive integer, let $n \cdot 1_F = 1_F + \cdots + 1_F$ (n times). Then two possibilities arise: either all the elements $n \cdot 1_F$ are distinct, or else $n \cdot 1_F = 0$ for some positive integer n.

Definition. The *characteristic* of a field F, denoted $\mathrm{ch}(F)$, is defined to be the smallest positive integer p such that $p \cdot 1_F = 0$ if such a p exists and is defined to be 0 otherwise.

It is easy to see that

$$\begin{aligned} n \cdot 1_F + m \cdot 1_F &= (m+n) \cdot 1_F \qquad \text{and that} \\ (n \cdot 1_F)(m \cdot 1_F) &= mn \cdot 1_F \end{aligned} \tag{13.1}$$

for positive integers m and n. It follows that the characteristic of a field is either 0 or a *prime* p (hence the choice of p in the definition above), since if $n = ab$ is composite with $n \cdot 1_F = 0$, then $ab \cdot 1_F = (a \cdot 1_F)(b \cdot 1_F) = 0$ and since F is a field, one of $a \cdot 1_F$ or $b \cdot 1_F$ is 0, so the smallest such integer is necessarily a prime. It also follows that if $n \cdot 1_F = 0$, then n is divisible by p.

Proposition 1. The characteristic of a field F, $\mathrm{ch}(F)$, is either 0 or a prime p. If $\mathrm{ch}(F) = p$ then for any $\alpha \in F$,

$$p \cdot \alpha = \underbrace{\alpha + \alpha + \cdots + \alpha}_{p \text{ times}} = 0.$$

Proof: Only the second statement has not been proved, and this follows immediately from the evident equality $p \cdot \alpha = p \cdot (1_F\alpha) = (p \cdot 1_F)(\alpha)$ in F.

Remark: This notion of a characteristic makes sense also for any integral domain and its characteristic will be the same as for its field of fractions.

Examples

(1) The fields $\mathbb{Q}$ and $\mathbb{R}$ both have characteristic 0: $\mathrm{ch}(\mathbb{Q}) = \mathrm{ch}(\mathbb{R}) = 0$. The integral domain $\mathbb{Z}$ also has characteristic 0.

(2) The (finite) field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ has characteristic p for any prime p.

(3) The integral domain $\mathbb{F}_p[x]$ of polynomials in the variable x with coefficients in the field $\mathbb{F}_p$ has characteristic p, as does its field of fractions $\mathbb{F}_p(x)$ (the field of rational functions in x with coefficients in $\mathbb{F}_p$).

If we define $(-n) \cdot 1_F = -(n \cdot 1_F)$ for positive n and $0 \cdot 1_F = 0$, then we have a natural ring homomorphism (by equation (1))

$$\begin{aligned} \varphi : \mathbb{Z} &\longrightarrow F \\ n &\longmapsto n \cdot 1_F \end{aligned}$$

and we can interpret the characteristic of F by noting that $\ker(\varphi) = \mathrm{ch}(F)\mathbb{Z}$. Taking the quotient by the kernel gives us an *injection* of either $\mathbb{Z}$ or $\mathbb{Z}/p\mathbb{Z}$ into F (depending on whether $\mathrm{ch}(F) = 0$ or $\mathrm{ch}(F) = p$). Since F is a field, we see that F contains a subfield isomorphic either to $\mathbb{Q}$ (the field of fractions of $\mathbb{Z}$) or to $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ (the field of fractions of $\mathbb{Z}/p\mathbb{Z}$) depending on the characteristic of F, and in either case is the smallest subfield of F containing 1_F (the field *generated* by 1_F in F).

Definition. The *prime subfield* of a field F is the subfield of F generated by the multiplicative identity 1_F of F. It is (isomorphic to) either $\mathbb{Q}$ (if $\mathrm{ch}(F) = 0$) or $\mathbb{F}_p$ (if $\mathrm{ch}(F) = p$).

Remark: We shall usually denote the identity 1_F of a field F simply by 1. Then in a field of characteristic p, one has $p \cdot 1 = 0$, frequently written simply $p = 0$ (for example, $2 = 0$ in a field of characteristic 2). It should be kept in mind, however, that this is a shorthand statement — the element "p" is really $p \cdot 1_F$ and is not a distinct element in F. This notation is useful in light of the second statement in Proposition 1.

Examples

(1) The prime subfield of both $\mathbb{Q}$ and $\mathbb{R}$ is $\mathbb{Q}$.

(2) The prime subfield of the field $\mathbb{F}_p(x)$ is isomorphic to $\mathbb{F}_p$, given by the constant polynomials.

Definition. If K is a field containing the subfield F, then K is said to be an *extension field* (or simply an *extension*) of F, denoted K/F or by the diagram

$$\begin{tikzcd} K \arrow[d, dash] \\ F \end{tikzcd}$$

In particular, every field F is an extension of its prime subfield. The field F is sometimes called the *base field* of the extension.

The notation K/F for a field extension is a shorthand for "K over F" and is not the quotient of K by F.

If K/F is any extension of fields, then the multiplication defined in K makes K into a *vector space* over F. In particular every field F can be considered as a vector space over its prime field.

Definition. The *degree* (or *relative degree* or *index*) of a field extension K/F, denoted $[K : F]$, is the dimension of K as a vector space over F (i.e., $[K : F] = \dim_F K$). The extension is said to be *finite* if $[K : F]$ is finite and is said to be *infinite* otherwise.

An important class of field extensions are those obtained by trying to solve equations over a given field F. For example, if $F = \mathbb{R}$ is the field of real numbers, then the simple equation $x^2 + 1 = 0$ does not have a solution in F. The question arises whether there is some larger field containing $\mathbb{R}$ in which this equation does have a solution, and it was this question that led Gauss to introduce the *complex numbers* $\mathbb{C} = \mathbb{R} + \mathbb{R}i$, where i is defined so that $i^2 + 1 = 0$. One then defines addition and multiplication in $\mathbb{C}$ by the usual rules familiar from elementary algebra and checks that in fact $\mathbb{C}$ so defined is a *field*, i.e., it is possible to find an inverse for every nonzero element of $\mathbb{C}$.

Given any field F and any polynomial $p(x) \in F[x]$ one can ask a similar question: does there exist an extension K of F containing a solution of the equation $p(x) = 0$ (i.e., containing a *root* of $p(x)$)? Note that we may assume here that the polynomial $p(x)$ is irreducible in $F[x]$ since a root of any factor of $p(x)$ is certainly a root of $p(x)$ itself. The answer is yes and follows almost immediately from our work on the polynomial ring $F[x]$. We first recall the following useful result on homomorphisms of fields (Corollary 10 of Chapter 7) which follows from the fact that the only ideals of a field F are 0 and F.

Proposition 2. Let $\varphi : F \to F'$ be a homomorphism of fields. Then φ is either identically 0 or is injective, so that the image of φ is either 0 or isomorphic to F.

Theorem 3. Let F be a field and let $p(x) \in F[x]$ be an irreducible polynomial. Then there exists a field K containing an isomorphic copy of F in which $p(x)$ has a root. Identifying F with this isomorphic copy shows that there exists an extension of F in which $p(x)$ has a root.

Proof: Consider the quotient

$$K = F[x]/(p(x))$$

of the polynomial ring $F[x]$ by the ideal generated by $p(x)$. Since by assumption $p(x)$ is an irreducible polynomial in the P.I.D. $F[x]$, the ideal $(p(x))$ is a *maximal* ideal. Hence K is actually a *field* (this is Proposition 12 of Chapter 7). The canonical projection π of $F[x]$ to the quotient $F[x]/(p(x))$ restricted to $F \subset F[x]$ gives a homomorphism $\varphi = \pi|_F : F \to K$ which is not identically 0 since it maps the identity 1 of F to the identity 1 of K. Hence by the proposition above, $\varphi(F) \cong F$ is an isomorphic copy of F contained in K. We identify F with its isomorphic image in K **and view** $\boldsymbol{F}$ **as a** *subfield* of K. If $\bar{x} = \pi(x)$ denotes the image of x in the quotient K, then

$$\begin{aligned} p(\bar{x}) &= \overline{p(x)} && \text{(since } \pi \text{ is a homomorphism)} \\ &= p(x) \pmod{p(x)} && \text{in } F[x]/(p(x)) \\ &= 0 && \text{in } F[x]/(p(x)) \end{aligned}$$

so that K does indeed contain a root of the polynomial $p(x)$. Then K is an extension of F in which the polynomial $p(x)$ has a root.

We shall use this result later to construct extensions of F containing *all* the roots of $p(x)$ (this is the notion of a *splitting field* and one of the central objects of interest in Galois theory).

To understand the field $K = F[x]/(p(x))$ constructed above more fully, it is useful to have a simple representation for the elements of this field. Since F is a subfield of K, we might in particular ask for a basis for K as a vector space over F.

Theorem 4. Let $p(x) \in F[x]$ be an irreducible polynomial of degree n over the field F and let K be the field $F[x]/(p(x))$. Let $\theta = x \bmod (p(x)) \in K$. Then the elements

$$1, \theta, \theta^2, \ldots, \theta^{n-1}$$

are a basis for K as a vector space over F, so the degree of the extension is n, i.e., $[K : F] = n$. Hence

$$K = \{a_0 + a_1\theta + a_2\theta^2 + \cdots + a_{n-1}\theta^{n-1} \mid a_0, a_1, \ldots, a_{n-1} \in F\}$$

consists of all polynomials of degree $< n$ in θ.

Proof: Let $a(x) \in F[x]$ be any polynomial with coefficients in F. Since $F[x]$ is a Euclidean Domain (this is Theorem 3 of Chapter 9), we may divide $a(x)$ by $p(x)$:

$$a(x) = q(x)p(x) + r(x) \qquad q(x), r(x) \in F[x] \text{ with } \deg r(x) < n.$$

Since $q(x)p(x)$ lies in the ideal $(p(x))$, it follows that $a(x) \equiv r(x) \bmod (p(x))$, which shows that every residue class in $F[x]/(p(x))$ is represented by a polynomial of degree less than n. Hence the images $1, \theta, \theta^2, \ldots, \theta^{n-1}$ of $1, x, x^2, \ldots, x^{n-1}$ in the quotient *span* the quotient as a vector space over F. It remains to see that these elements are linearly independent, so form a *basis* for the quotient over F.

If the elements $1, \theta, \theta^2, \ldots, \theta^{n-1}$ were not linearly independent in K, then there would be a linear combination

$$b_0 + b_1\theta + b_2\theta^2 + \cdots + b_{n-1}\theta^{n-1} = 0$$

in K, with $b_0, b_1, \ldots, b_{n-1} \in F$, not all 0. This is equivalent to

$$b_0 + b_1x + b_2x^2 + \cdots + b_{n-1}x^{n-1} \equiv 0 \bmod (p(x))$$

i.e.,

$$p(x) \text{ divides } b_0 + b_1x + b_2x^2 + \cdots + b_{n-1}x^{n-1}$$

in $F[x]$. But this is impossible, since $p(x)$ is of degree n and the degree of the nonzero polynomial on the right is $< n$. This proves that $1, \theta, \theta^2, \dots, \theta^{n-1}$ are a basis for K over F, so that $[K : F] = n$ by definition. The last statement of the theorem is clear.

This theorem provides an easy description of the elements of the field $F[x]/(p(x))$ as polynomials of degree $< n$ in θ where θ is an element (in K) with $p(\theta) = 0$. It remains only to see how to add and multiply elements written in this form. The addition in the quotient $F[x]/(p(x))$ is just usual addition of polynomials. The multiplication of polynomials $a(x)$ and $b(x)$ in the quotient $F[x]/(p(x))$ is performed by finding the product $a(x)b(x)$ in $F[x]$, then finding the representative of degree $< n$ for the coset $a(x)b(x) + (p(x))$ (as in the proof above) by dividing $a(x)b(x)$ by $p(x)$ and finding the remainder.

This can also be done easily in terms of θ as follows: We may suppose $p(x)$ is monic (since its roots and the ideal it generates do not change by multiplying by a constant), say $p(x) = x^n + p_{n-1}x^{n-1} + \cdots + p_1x + p_0$. Then in K, since $p(\theta) = 0$, we have

$$\theta^n = -(p_{n-1}\theta^{n-1} + \cdots + p_1\theta + p_0)$$

i.e., θ^n is a linear combination of lower powers of θ. Multiplying both sides by θ and replacing the θ^n on the right hand side by these lower powers again, we see that also θ^{n+1} is a polynomial of degree $< n$ in θ. Similarly, any positive power of θ can be written as a polynomial of degree $< n$ in θ, hence *any* polynomial in θ can be written as a polynomial of degree $< n$ in θ. Multiplication in K is now easily performed: one simply writes the product of two polynomials of degree $< n$ in θ as another polynomial of degree $< n$ in θ.

We summarize this as:

Corollary 5. Let K be as in Theorem 4, and let $a(\theta), b(\theta) \in K$ be two polynomials of degree $< n$ in θ. Then addition in K is defined simply by usual polynomial addition and multiplication in K is defined by

$$a(\theta)b(\theta) = r(\theta)$$

where $r(x)$ is the remainder (of degree $< n$) obtained after dividing the polynomial $a(x)b(x)$ by $p(x)$ in $F[x]$.

By the results proved above, this definition of addition and multiplication on the polynomials of degree $< n$ in θ make K into a *field*, so that one can also *divide* by nonzero elements as well, which is not so immediately obvious from the definitions of the operations.

It is also important in Theorem 4 that the polynomial $p(x)$ be *irreducible* over F. In general the addition and multiplication in Corollary 5 (which can be defined in the same way for any polynomial $p(x)$) do *not* make the polynomials of degree $< n$ in θ into a field if $p(x)$ is not irreducible. In fact, this set is not even an integral domain in general (its structure is given by Proposition 16 of Chapter 9). To describe the *field* containing a root θ of a general polynomial $f(x)$ over F, $f(x)$ is factored into irreducibles in $F[x]$ and the results above are applied to an irreducible factor $p(x)$ of $f(x)$ having θ as a root. We shall consider this more in the following sections.

Examples

(1) If we apply this construction to the special case $F = \mathbb{R}$ and $p(x) = x^2 + 1$ then we obtain the field

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

which is an extension of degree 2 of $\mathbb{R}$ in which $x^2 + 1$ has a root. The elements of this field are of the form $a + b\theta$ for $a, b \in \mathbb{R}$. Addition is defined by

$$(a + b\theta) + (c + d\theta) = (a + c) + (b + d)\theta. \tag{13.2a}$$

To multiply we use the fact that $\theta^2 + 1 = 0$, i.e., $\theta^2 = -1$ in K. (Alternatively, note that -1 is also the remainder when x^2 is divided by $x^2 + 1$ in $\mathbb{R}[x]$.) Then

$$\begin{aligned}(a + b\theta)(c + d\theta) &= ac + (ad + bc)\theta + bd\theta^2 \\ &= ac + (ad + bc)\theta + bd(-1) \\ &= (ac - bd) + (ad + bc)\theta. \end{aligned} \tag{13.2b}$$

These are, up to changing θ to i, the formulas for adding and multiplying in $\mathbb{C}$. Put another way, the map

$$\begin{aligned}\varphi : \mathbb{R}[x]/(x^2 + 1) &\longrightarrow \mathbb{C} \\ a + bx &\mapsto a + bi\end{aligned}$$

is a homomorphism. Since it is bijective (as a map of vector spaces over the reals, for example), it is an isomorphism. Notice that instead of taking the existence of $\mathbb{C}$ for granted (along with the fairly tedious verification that it is in fact a field), we could have *defined* $\mathbb{C}$ by this isomorphism. Then the fact that it is a field is a consequence of Theorem 4.

(2) Take now $F = \mathbb{Q}$ to be the field of rational numbers and again take $p(x) = x^2 + 1$ (still irreducible over $\mathbb{Q}$, of course). Then the same construction, with the same addition and multiplication formulas as (2a) and (2b) above, except that now a and b are elements of $\mathbb{Q}$, defines a field extension $\mathbb{Q}(i)$ of $\mathbb{Q}$ of degree 2 containing a root i of $x^2 + 1$.

(3) Take $F = \mathbb{Q}$ and $p(x) = x^2 - 2$, irreducible over $\mathbb{Q}$ by Eisenstein's Criterion, for example. Then we obtain a field extension of $\mathbb{Q}$ of degree 2 containing a square root θ of 2, denoted $\mathbb{Q}(\theta)$. If we denote θ by $\sqrt{2}$, the elements of this field are of the form

$$a + b\sqrt{2}, \qquad a, b \in \mathbb{Q}$$

with addition defined by

$$(a + b\sqrt{2}) + (c + d\sqrt{2}) = (a + c) + (b + d)\sqrt{2}$$

and multiplication defined by

$$(a + b\sqrt{2})(c + d\sqrt{2}) = (ac + 2bd) + (ad + bc)\sqrt{2}.$$

(4) Let $F = \mathbb{Q}$ and $p(x) = x^3 - 2$, irreducible again by Eisenstein. Denoting a root of $p(x)$ by θ, we obtain the field

$$\mathbb{Q}[x]/(x^3 - 2) \cong \{a + b\theta + c\theta^2 \mid a, b, c \in \mathbb{Q}\}$$

with $\theta^3 = 2$, an extension of degree 3. To find the inverse of, say, $1 + \theta$ in this field, we can proceed as follows: By the Euclidean Algorithm in $\mathbb{Q}[x]$ there are polynomials $a(x)$ and $b(x)$ with

$$a(x)(1 + x) + b(x)(x^3 - 2) = 1$$

(since $p(x) = x^3 - 2$ is irreducible, it is relatively prime to every polynomial of smaller degree). In the quotient field this equation implies that $a(\theta)$ is the inverse of $1 + \theta$. In this case, a simple computation shows that we can take $a(x) = \frac{1}{3}(x^2 - x + 1)$ (and $b(x) = -\frac{1}{3}$), so that

$$(1 + \theta)^{-1} = \frac{\theta^2 - \theta + 1}{3}.$$

(5) In general, if $\theta \in K$ is a root of the irreducible polynomial

$$p(x) = p_n x^n + p_{n-1} x^{n-1} + \cdots + p_1 x + p_0$$

we can compute $\theta^{-1} \in K$ from

$$\theta(p_n \theta^{n-1} + p_{n-1}\theta^{n-2} + \cdots + p_1) = -p_0$$

namely

$$\theta^{-1} = \frac{-1}{p_0}(p_n \theta^{n-1} + p_{n-1}\theta^{n-2} + \cdots + p_1) \in K$$

(note that $p_0 \neq 0$ since $p(x)$ is irreducible).

Remark: Determining inverses in extensions of this type may be familiar from elementary algebra in the case of $\mathbb{C}$ or Example 3 under the name "rationalizing denominators." The last two examples indicates a procedure which is much more general than the ad hoc procedures of elementary algebra.

(6) Take $F = \mathbb{F}_2$, the finite field with two elements, and $p(x) = x^2 + x + 1$, which we have previously checked is irreducible over $\mathbb{F}_2$. Here we obtain a degree 2 extension of $\mathbb{F}_2$

$$\mathbb{F}_2[x]/(x^2 + x + 1) \cong \{a + b\theta \mid a, b \in \mathbb{F}_2\}$$

where $\theta^2 = -\theta - 1 = \theta + 1$. Multiplication in this field $\mathbb{F}_2(\theta)$ (which contains four elements) is defined by

$$\begin{aligned}(a + b\theta)(c + d\theta) &= ac + (ad + bc)\theta + bd\theta^2 \\ &= ac + (ad + bc)\theta + bd(\theta + 1) \\ &= (ac + bd) + (ad + bc + bd)\theta.\end{aligned}$$

(7) Let $F = k(t)$ be the field of rational functions in the variable t over a field k (for example, $k = \mathbb{Q}$ or $k = \mathbb{F}_p$). Let $p(x) = x^2 - t \in F[x]$. Then $p(x)$ is irreducible (it is Eisenstein at the prime (t) in $k[t]$). If we denote a root by θ, the corresponding degree 2 field extension $F(\theta)$ consists of the elements

$$\{a(t) + b(t)\theta \mid a(t), b(t) \in F\}$$

where the coefficients $a(t)$ and $b(t)$ are rational functions in t with coefficients in k and where $\theta^2 = t$.

Suppose F is a subfield of a field K and $\alpha \in K$ is an element of K. Then the collection of subfields of K containing both F and α is nonempty (K is such a field, for example). Since the intersection of subfields is again a subfield, it follows that there is a unique minimal subfield of K containing both F and α (the intersection of all subfields with this property). Similar remarks apply if α is replaced by a collection $\alpha, \beta, \ldots$ of elements of K.

Definition. Let K be an extension of the field F and let $\alpha, \beta, \cdots \in K$ be a collection of elements of K. Then the smallest subfield of K containing both F and the elements $\alpha, \beta, \ldots$, denoted $F(\alpha, \beta, \ldots)$ is called the field *generated by* $\alpha, \beta, \ldots$ *over* F.

Definition. If the field K is generated by a single element α over F, $K = F(\alpha)$, then K is said to be a *simple* extension of F and the element α is called a *primitive element* for the extension.

We shall later characterize which extensions of a field F are simple. In particular we shall prove that every finite extension of a field of characteristic 0 is a simple extension.

The connection between the simple extension $F(\alpha)$ generated by α over F where α is a root of some irreducible polynomial $p(x)$ and the field constructed in Theorem 3 is provided by the following:

Theorem 6. Let F be a field and let $p(x) \in F[x]$ be an irreducible polynomial. Suppose K is an extension field of F containing a root α of $p(x)$: $p(\alpha) = 0$. Let $F(\alpha)$ denote the subfield of K generated over F by α. Then

$$F(\alpha) \cong F[x]/(p(x)).$$

Remark: This theorem says that *any* field over F in which $p(x)$ contains a root contains a subfield isomorphic to the extension of F constructed in Theorem 3 and that this field is (up to isomorphism) the smallest extension of F containing such a root. The difference between this result and Theorem 3 is that Theorem 6 *assumes* the existence of a root α of $p(x)$ in some field K and the major point of Theorem 3 is *proving* that there exists such an extension field K.

Proof: There is a natural homomorphism

$$\begin{aligned} \varphi : F[x] &\longrightarrow F(\alpha) \subseteq K \\ a(x) &\longmapsto a(\alpha) \end{aligned}$$

obtained by mapping F to F by the identity map and sending x to α and then extending so that the map is a ring homomorphism (i.e., the polynomial $a(x)$ in x maps to the polynomial $a(\alpha)$ in α). Since $p(\alpha) = 0$ by assumption, the element $p(x)$ is in the kernel of φ, so we obtain an induced homomorphism (also denoted φ):

$$\varphi : F[x]/(p(x)) \longrightarrow F(\alpha).$$

But since $p(x)$ is irreducible, the quotient on the left is a *field*, and φ is not the 0 map (it is the identity on F, for example), hence φ is an isomorphism of the field on the left with its image. Since this image is then a subfield of $F(\alpha)$ containing F and containing α, by the definition of $F(\alpha)$ the map must be surjective, proving the theorem.

Combined with Corollary 5, this determines the field $F(\alpha)$ when α is a root of an irreducible polynomial $p(x)$:

Corollary 7. Suppose in Theorem 6 that $p(x)$ is of degree n. Then

$$F(\alpha) = \{a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_{n-1}\alpha^{n-1} \mid a_0, a_1, \ldots, a_{n-1} \in F\} \subseteq K.$$

Describing fields generated by more than one element is more complicated and we shall return to this question in the following section.

Examples

(1) In Example 3 above, we have determined the field $\mathbb{Q}(\sqrt{2})$ generated over $\mathbb{Q}$ by the element $\sqrt{2} \in \mathbb{R}$, having suggestively denoted the abstract solution θ of the equation $x^2 - 2 = 0$ by the symbol $\sqrt{2}$, which has an independent meaning in the field $\mathbb{R}$ (namely the *positive* square root of 2 in $\mathbb{R}$).

(2) The equation $x^2 - 2 = 0$ has another solution in $\mathbb{R}$, namely $-\sqrt{2}$, the *negative* square root of 2 in $\mathbb{R}$. The field generated over $\mathbb{Q}$ by this solution consists of the elements $\{a + b(-\sqrt{2}) \mid a, b \in \mathbb{Q}\}$, and is again isomorphic to the field in Example 3 above (hence also isomorphic to the field just considered, the isomorphism given explicitly by $a + b\sqrt{2} \mapsto a - b\sqrt{2}$). As a subset of $\mathbb{R}$ this is the same set of elements as in Example 1.

(3) Similarly, if we use the symbol $\sqrt[3]{2}$ to denote the (positive) cube root of 2 in $\mathbb{R}$, then the field generated by $\sqrt[3]{2}$ over $\mathbb{Q}$ in $\mathbb{R}$ consists of the elements

$$\{a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 \mid a, b, c \in \mathbb{Q}\}$$

and is isomorphic to the field constructed in Example 4 above.

(4) The equation $x^3 - 2 = 0$ has no further solutions in $\mathbb{R}$, but there are two additional solutions in $\mathbb{C}$ given by $\sqrt[3]{2}(\frac{-1 + i\sqrt{3}}{2})$ and $\sqrt[3]{2}(\frac{-1 - i\sqrt{3}}{2})$ ($\sqrt{3}$ denoting the positive real square root of 3) as can easily be checked. The fields generated by either of these two elements over $\mathbb{Q}$ are subfields of $\mathbb{C}$ (but not of $\mathbb{R}$) and are both isomorphic to the field constructed in the previous example (and to Example 4 earlier).

As Theorem 6 indicates, the roots of an irreducible polynomial $p(x)$ are *algebraically indistinguishable* in the sense that the fields obtained by adjoining any root of an irreducible polynomial are isomorphic. In the last two examples above, the fields obtained by adjoining one of the three possible (complex) roots of $x^3 - 2 = 0$ to $\mathbb{Q}$ were all algebraically isomorphic. The fields were distinguished not by their algebraic properties, but by whether their elements were *real*, which involves *continuous* operations.

The fact that different roots of the same irreducible polynomial have the same algebraic properties can be extended slightly, as follows:

Let $\varphi : F \xrightarrow{\sim} F'$ be an isomorphism of fields. The map φ induces a ring isomorphism (also denoted φ)

$$\varphi : F[x] \xrightarrow{\sim} F'[x]$$

defined by applying φ to the coefficients of a polynomial in $F[x]$. Let $p(x) \in F[x]$ be an irreducible polynomial and let $p'(x) \in F'[x]$ be the polynomial obtained by applying the map φ to the coefficients of $p(x)$, i.e., the image of $p(x)$ under φ. The isomorphism φ maps the maximal ideal $(p(x))$ to the ideal $(p'(x))$, so this ideal is also maximal, which shows that $p'(x)$ is also irreducible in $F'[x]$. The following theorem shows that the fields obtained by adjoining a root of $p(x)$ to F and a root of $p'(x)$ to F' have the same algebraic structure (i.e., are isomorphic):

Theorem 8. Let $\varphi : F \xrightarrow{\sim} F'$ be an isomorphism of fields. Let $p(x) \in F[x]$ be an irreducible polynomial and let $p'(x) \in F'[x]$ be the irreducible polynomial obtained by applying the map φ to the coefficients of $p(x)$. Let α be a root of $p(x)$ (in some extension of F) and let β be a root of $p'(x)$ (in some extension of F'). Then there is an isomorphism

$$\begin{aligned}\sigma : F(\alpha) &\xrightarrow{\sim} F'(\beta) \\ \alpha &\longmapsto \beta\end{aligned}$$

mapping α to β and extending φ, i.e., such that σ restricted to F is the isomorphism φ.

Proof: As noted above, the isomorphism φ induces a natural isomorphism from $F[x]$ to $F'[x]$ which maps the maximal ideal $(p(x))$ to the maximal ideal $(p'(x))$. Taking the quotients by these ideals, we obtain an isomorphism of fields

$$F[x]/(p(x)) \xrightarrow{\sim} F'[x]/(p'(x)).$$

By Theorem 6 the field on the left is isomorphic to $F(\alpha)$ and by the same theorem the field on the right is isomorphic to $F'(\beta)$. Composing these isomorphisms, we obtain the isomorphism σ. It is clear that the restriction of this isomorphism to F is φ, completing the proof.

This extension theorem will be of considerable use when we consider Galois Theory later. It can be represented pictorially by the diagram

$$\begin{tikzcd}
\sigma : & F(\alpha) \arrow[r, "\sim"] \arrow[d, dash] & F'(\beta) \arrow[d, dash] \\
\varphi : & F \arrow[r, "\sim"] & F'
\end{tikzcd}$$

EXERCISES

1. Show that $p(x) = x^3 + 9x + 6$ is irreducible in $\mathbb{Q}[x]$. Let θ be a root of $p(x)$. Find the inverse of $1 + \theta$ in $\mathbb{Q}(\theta)$.
2. Show that $x^3 - 2x - 2$ is irreducible over $\mathbb{Q}$ and let θ be a root. Compute $(1+\theta)(1+\theta+\theta^2)$ and $\dfrac{1+\theta}{1+\theta+\theta^2}$ in $\mathbb{Q}(\theta)$.
3. Show that $x^3 + x + 1$ is irreducible over $\mathbb{F}_2$ and let θ be a root. Compute the powers of θ in $\mathbb{F}_2(\theta)$.
4. Prove directly that the map $a + b\sqrt{2} \mapsto a - b\sqrt{2}$ is an isomorphism of $\mathbb{Q}(\sqrt{2})$ with itself.
5. Suppose α is a rational root of a monic polynomial in $\mathbb{Z}[x]$. Prove that α is an integer.
6. Show that if α is a root of $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ then $a_n\alpha$ is a root of the monic polynomial $x^n + a_{n-1}x^{n-1} + a_n a_{n-2} x^{n-2} + \cdots + a_n^{n-2} a_1 x + a_n^{n-1} a_0$.
7. Prove that $x^3 - nx + 2$ is irreducible for $n \neq -1, 3, 5$.
8. Prove that $x^5 - ax - 1 \in \mathbb{Z}[x]$ is irreducible unless $a = 0, 2$ or -1. The first two correspond to linear factors, the third corresponds to the factorization $(x^2 - x + 1)(x^3 + x^2 - 1)$.

13.2 ALGEBRAIC EXTENSIONS

Let F be a field and let K be an extension of F.

Definition. The element $\alpha \in K$ is said to be *algebraic* over F if α is a root of some nonzero polynomial $f(x) \in F[x]$. If α is not algebraic over F (i.e., is not the root of any nonzero polynomial with coefficients in F) then α is said to be *transcendental* over F. The extension K/F is said to be *algebraic* if every element of K is algebraic over F.

Note that if α is algebraic over a field F then it is algebraic over any extension field L of F (if $f(x)$ having α as a root has coefficients in F then it also has coefficients in L).

Proposition 9. Let α be algebraic over F. Then there is a unique monic irreducible polynomial $m_{\alpha,F}(x) \in F[x]$ which has α as a root. A polynomial $f(x) \in F[x]$ has α as a root if and only if $m_{\alpha,F}(x)$ divides $f(x)$ in $F[x]$.

Proof: Let $g(x) \in F[x]$ be a polynomial of minimal degree having α as a root. Multiplying $g(x)$ by a constant, we may assume $g(x)$ is monic. Suppose $g(x)$ were reducible in $F[x]$, say $g(x) = a(x)b(x)$ with $a(x), b(x) \in F[x]$ both of degree smaller than the degree of $g(x)$. Then $g(\alpha) = a(\alpha)b(\alpha)$ in K, and since K is a field, either $a(\alpha) = 0$ or $b(\alpha) = 0$, contradicting the minimality of the degree of $g(x)$. It follows that $g(x)$ is a monic irreducible polynomial having α as a root. Suppose now that $f(x) \in F[x]$ is any polynomial having α as a root. By the Euclidean Algorithm in $F[x]$ there are polynomials $q(x), r(x) \in F[x]$ such that

$$f(x) = q(x)g(x) + r(x) \qquad \text{with} \quad \deg r(x) < \deg g(x).$$

Then $f(\alpha) = q(\alpha)g(\alpha) + r(\alpha)$ in K and since α is a root of both $f(x)$ and $g(x)$, we obtain $r(\alpha) = 0$, which contradicts the minimality of $g(x)$ unless $r(x) = 0$. Hence $g(x)$ divides any polynomial $f(x)$ in $F[x]$ having α as a root and, in particular, would divide any other monic irreducible polynomial in $F[x]$ having α as a root. This proves that $m_{\alpha,F}(x) = g(x)$ is unique and completes the proof of the proposition.

Corollary 10. If L/F is an extension of fields and α is algebraic over both F and L, then $m_{\alpha,L}(x)$ divides $m_{\alpha,F}(x)$ in $L[x]$.

Proof: This is immediate from the second statement in Proposition 9 applied to L, since $m_{\alpha,F}(x)$ is a polynomial in $L[x]$ having α as a root.

Definition. The polynomial $m_{\alpha,F}(x)$ (or just $m_\alpha(x)$ if the field F is understood) in Proposition 9 is called the *minimal polynomial* for α over F. The *degree* of $m_\alpha(x)$ is called the *degree* of α.

Note that by the proposition, a monic polynomial over F with α as a root is the minimal polynomial for α over F if and only if it is irreducible over F. Exercise 20 gives one method for computing the minimal polynomial for α over F, and the theory of Gröbner bases can be used to compute the minimal polynomial for other elements in $F(\alpha)$ (cf. Proposition 10 and Exercise 48 in Section 15.1).

Proposition 11. Let α be algebraic over the field F and let $F(\alpha)$ be the field generated by α over F. Then

$$F(\alpha) \cong F[x]/(m_\alpha(x))$$

so that in particular

$$[F(\alpha) : F] = \deg\ m_\alpha(x) = \deg \alpha,$$

i.e., the degree of α over F is the degree of the extension it generates over F.

Proof: This follows immediately from Theorem 6.

Examples

(1) The minimal polynomial for $\sqrt{2}$ over $\mathbb{Q}$ is $x^2 - 2$ and $\sqrt{2}$ is of degree 2 over $\mathbb{Q}$: $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$.

(2) The minimal polynomial for $\sqrt[3]{2}$ over $\mathbb{Q}$ is $x^3 - 2$ and $\sqrt[3]{2}$ is of degree 3 over $\mathbb{Q}$: $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$.

(3) Similarly, for any $n > 1$, the polynomial $x^n - 2$ is irreducible over $\mathbb{Q}$ since it is Eisenstein. Denoting a root of this polynomial by $\sqrt[n]{2}$ (where as usual we reserve this symbol to denote the *positive* n^{th} root of 2 if we want to view this root as an element of $\mathbb{R}$, and where the symbol denotes any one of the algebraically indistinguishable abstract solutions in general), we have $[\mathbb{Q}(\sqrt[n]{2}) : \mathbb{Q}] = n$.

(4) The minimal polynomial and the degree of an element α depend on the base field. For example, over $\mathbb{R}$, the element $\sqrt[n]{2}$ is of degree *one*, with minimal polynomial $m_{\sqrt[n]{2},\mathbb{R}}(x) = x - \sqrt[n]{2}$.

(5) Consider the polynomial $p(x) = x^3 - 3x - 1$ over $\mathbb{Q}$, which is irreducible over $\mathbb{Q}$ since it is a cubic which has no rational root (cf. Proposition 11 of Chapter 9). Hence $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$ for any root α of $p(x)$. For future reference we note that a quick sketch of the graph of this function over the real numbers shows that the graph crosses the x-axis precisely once in the interval [0,2], i.e., there is precisely one real number α, $0 < \alpha < 2$ satisfying $\alpha^3 - 3\alpha - 1 = 0$.

Proposition 12. The element α is algebraic over F if and only if the simple extension $F(\alpha)/F$ is finite. More precisely, if α is an element of an extension of degree n over F then α satisfies a polynomial of degree at most n over F and if α satisfies a polynomial of degree n over F then the degree of $F(\alpha)$ over F is at most n.

Proof: If α is algebraic over F, then the degree of the extension $F(\alpha)/F$ is the degree of the minimal polynomial for α over F. Hence the extension is finite, of degree $\leq n$ if α satisfies a polynomial of degree n. Conversely, suppose α is an element of an extension of degree n over F (for example, if $[F(\alpha) : F] = n$). Then the $n + 1$ elements

$$1, \alpha, \alpha^2, \ldots, \alpha^n$$

of $F(\alpha)$ are linearly dependent over F, say

$$b_0 + b_1\alpha + b_2\alpha^2 + \cdots + b_n\alpha^n = 0$$

with $b_0, b_1, b_2, \dots, b_n \in F$ not all 0. Hence α is the root of a nonzero polynomial with coefficients in F (of degree $\leq n$), which proves α is algebraic over F and also proves the second statement of the proposition.

Corollary 13. If the extension K/F is finite, then it is algebraic.

Proof: If $\alpha \in K$, then the subfield $F(\alpha)$ is in particular a subspace of the vector space K over F. Hence $[F(\alpha) : F] \leq [K : F]$ and so α is algebraic over F by the proposition.

Remark: We shall prove below a sort of converse to this result (Theorem 17), but note that there are infinite algebraic extensions (we shall have an example later), so the literal converse of this corollary is not true.

Example: (Quadratic Extensions over Fields of Characteristic $\neq 2$)

Let F be a field of characteristic $\neq 2$ (for example, any field of characteristic 0, such as $\mathbb{Q}$) and let K be an extension of F of degree 2, $[K : F] = 2$. Let α be any element of K not contained in F. By the proposition above, α satisfies an equation of degree at most 2 over F. This equation cannot be of degree 1, since α is not an element of F by assumption. It follows that the minimal polynomial of α is a monic quadratic

$$m_\alpha(x) = x^2 + bx + c \qquad b, c \in F.$$

Since $F \subset F(\alpha) \subseteq K$ and $F(\alpha)$ is already a vector space over F of dimension 2, we have $K = F(\alpha)$.

The roots of this quadratic equation can be determined by the quadratic formula, which is valid over any field of characteristic $\neq 2$ (the formula is obtained as in elementary algebra by completing the square):

$$\alpha = \frac{-b \pm \sqrt{b^2 - 4c}}{2}$$

(the reason for requiring the characteristic of F not be 2 is that we must divide by 2). Here $b^2 - 4c$ is not a square in F since α is not an element of F and the symbol $\sqrt{b^2 - 4c}$ denotes a root of the equation $x^2 - (b^2 - 4c) = 0$ in K (see the end of the next paragraph). Note that here there is no natural choice of one of the roots analogous to choosing the *positive* square root of 2 in $\mathbb{R}$ — the roots are algebraically indistinguishable.

Now $F(\alpha) = F(\sqrt{b^2 - 4c})$ as follows: by the formula above, α is an element of the field on the right, hence $F(\alpha) \subseteq F(\sqrt{b^2 - 4c})$. Conversely, $\sqrt{b^2 - 4c} = \mp(b + 2\alpha)$ shows that $\sqrt{b^2 - 4c}$ is an element of $F(\alpha)$, which gives the reverse inclusion $F(\sqrt{b^2 - 4c}) \subseteq F(\alpha)$ (and incidentally shows that the equation $x^2 - (b^2 - 4c) = 0$ does have a solution in K).

It follows that any extension K of F of degree 2 is of the form $F(\sqrt{D})$ where D is an element of F which is not a square in F, and conversely, every such extension is an extension of degree 2 of F. For this reason, extensions of degree 2 of a field F are called *quadratic* extensions of F.

Suppose that F is a subfield of a field K which in turn is a subfield of a field L. Then there are three associated extension degrees — the dimension of K and L as vector spaces over F, and the dimension of L as a vector space over K.

Theorem 14. Let $F \subseteq K \subseteq L$ be fields. Then

$$[L : F] = [L : K][K : F],$$

i.e. extension degrees are multiplicative, where if one side of the equation is infinite, the other side is also infinite. Pictorially,

$$\overbrace{\underbrace{F \quad \subseteq}_{[K:F]} \quad K \quad \underbrace{\subseteq \quad L}_{[L:K]}}^{[L:F]}$$

Proof: Suppose first that $[L : K] = m$ and $[K : F] = n$ are finite. Let $\alpha_1, \alpha_2, \dots, \alpha_m$ be a basis for L over K and let $\beta_1, \beta_2, \dots, \beta_n$ be a basis for K over F. Then every element of L can be written as a linear combination

$$a_1\alpha_1 + a_2\alpha_2 + \cdots + a_m\alpha_m$$

where $a_1, \dots, a_m$ are elements of K, hence are F-linear combinations of $\beta_1, \dots, \beta_n$:

$$a_i = b_{i1}\beta_1 + b_{i2}\beta_2 + \cdots + b_{in}\beta_n \qquad i = 1, 2, \dots, m \tag{13.3}$$

where the b_{ij} are elements of F. Substituting these expressions in for the coefficients a_i above, we see that every element of L can be written as a linear combination

$$\sum_{\substack{i=1,2,\dots,m \\ j=1,2,\dots,n}} b_{ij}\alpha_i\beta_j$$

of the mn elements $\alpha_i\beta_j$ with coefficients in F. Hence these elements *span* L as a vector space over F.

Suppose now that we had a linear relation in L

$$\sum_{\substack{i=1,2,\dots,m \\ j=1,2,\dots,n}} b_{ij}\alpha_i\beta_j = 0$$

with coefficients b_{ij} in F. Then defining the elements $a_i \in K$ by equation (3) above, this linear relation could be written

$$a_1\alpha_1 + a_2\alpha_2 + \cdots + a_m\alpha_m = 0.$$

Since the α_i are a basis for L over K, it follows that all the coefficients $a_i, i = 1, 2, \dots, m$ must be 0, i.e., that

$$b_{i1}\beta_1 + b_{i2}\beta_2 + \cdots + b_{in}\beta_n = 0 \qquad i = 1, 2, \dots, m$$

in K. Since now the β_j, $j = 1, 2, \dots, n$ form a basis for K over F, this implies $b_{ij} = 0$ for all i and j. Hence the elements $\alpha_i\beta_j$ are linearly independent over F, so form a basis for L over F and $[L : F] = mn = [L : K][K : F]$, as claimed.

If $[K : F]$ is infinite, then there are infinitely many elements of K, hence of L, which are linearly independent over F, so that $[L : F]$ is also infinite. Similarly, if $[L : K]$ is infinite, there are infinitely many elements of L linearly independent over K, so certainly linearly independent over F, so again $[L : F]$ is infinite. Finally, if $[L : K]$ and $[K : F]$ are both finite, then the proof above shows $[L : F]$ is finite, so that $[L : F]$ infinite implies at least one of $[L : K]$ and $[K : F]$ is infinite, completing the proof.

Remark: Note the similarity of this result with the result on group orders proved in Part I. As with diagrams involving groups we shall frequently indicate the relative degrees of extensions in field diagrams.

The multiplicativity of extension degrees is extremely useful in computations. A particular application is the following:

Corollary 15. Suppose L/F is a finite extension and let K be any subfield of L containing F, $F \subseteq K \subseteq L$. Then $[K : F]$ divides $[L : F]$.

Proof: This is immediate.

Examples

(1) The element $\sqrt{2}$ is not contained in the field $\mathbb{Q}(\alpha)$ where α is the real root of $x^3 - 3x - 1$ between 0 and 2, since we have already determined that $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$ and $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$ and 2 does not divide 3. Note that it is not so easy to prove directly that $\sqrt{2}$ cannot be written as a rational linear combination of $1, \alpha, \alpha^2$.

(2) Let as usual $\sqrt[6]{2}$ denote the positive real 6^{th} root of 2. Then $[\mathbb{Q}(\sqrt[6]{2}) : \mathbb{Q}] = 6$. Since $(\sqrt[6]{2})^3 = \sqrt{2}$ we have $\mathbb{Q}(\sqrt{2}) \subset \mathbb{Q}(\sqrt[6]{2})$ and by the multiplicativity of extension degrees, $[\mathbb{Q}(\sqrt[6]{2}) : \mathbb{Q}(\sqrt{2})] = 3$. This gives us the field diagram

$$\overbrace{\underbrace{\mathbb{Q} \quad \subset \quad}_{2} \mathbb{Q}(\sqrt{2}) \underbrace{\quad \subset \quad \mathbb{Q}(\sqrt[6]{2})}_{3}}^{6}$$

In particular, this shows that the minimal polynomial for $\sqrt[6]{2}$ over $\mathbb{Q}(\sqrt{2})$ is of degree 3. It is therefore the polynomial $x^3 - \sqrt{2}$. Note that showing directly that this polynomial is irreducible over $\mathbb{Q}(\sqrt{2})$ is not completely trivial.

By Theorem 14 a finite extension of a finite extension is finite. The next results use this to show that an extension generated by a finite number of algebraic elements is finite (extending Proposition 12).

Definition. An extension K/F is *finitely generated* if there are elements $\alpha_1, \alpha_2, \dots, \alpha_k$ in K such that $K = F(\alpha_1, \alpha_2, \dots, \alpha_k)$.

Recall that the field generated over F by a collection of elements in a field K is the smallest subfield of K containing these elements and F. The next lemma will show that for finitely generated extensions this field can be obtained recursively by a series of simple extensions.

Lemma 16. $F(\alpha, \beta) = (F(\alpha))(\beta)$, i.e., the field generated over F by α and β is the field generated by β over the field $F(\alpha)$ generated by α.

Proof: This follows by the minimality of the fields in question. The field $F(\alpha, \beta)$ contains F and α, hence contains the field $F(\alpha)$, and since it also contains β, we have the inclusion $(F(\alpha))(\beta) \subseteq F(\alpha, \beta)$ by the minimality of the field $(F(\alpha))(\beta)$. Since the field $(F(\alpha))(\beta)$ contains F, α and β, by the minimality of $F(\alpha, \beta)$ we have the reverse inclusion $F(\alpha, \beta) \subseteq (F(\alpha))(\beta)$, which proves the lemma.

By the lemma we have

$$K = F(\alpha_1, \alpha_2, \dots, \alpha_k) = (F(\alpha_1, \alpha_2, \dots, \alpha_{k-1}))(\alpha_k)$$

and so by iterating, we see that K is obtained by taking the field F_1 generated over F by α_1, then the field F_2 generated *over* F_1 (this is important) by α_2, and so on, with $F_k = K$. This gives a sequence of fields:

$$F = F_0 \subseteq F_1 \subseteq F_2 \subseteq \cdots \subseteq F_k = K$$

where

$$F_{i+1} = F_i(\alpha_{i+1}) \qquad i = 0, 1, \dots, k-1.$$

Suppose now that the elements $\alpha_1, \alpha_2, \dots, \alpha_k$ are algebraic over F of degrees $n_1, n_2, \dots, n_k$ (so a priori are algebraic over any extension of F). Then the extensions in this sequence are simple extensions of the type considered in Proposition 11. The relative extension degree $[F_{i+1} : F_i]$ is equal to the degree of the minimal polynomial of α_{i+1} over F_i, which is at most n_{i+1} (and equals n_{i+1} if and only if the minimal polynomial of α_{i+1} over F remains irreducible over F_i). By the multiplicativity of extension degrees, we see that

$$[K : F] = [F_k : F_{k-1}][F_{k-1} : F_{k-2}] \cdots [F_1 : F_0]$$

is also finite, and $\leq n_1 n_2 \cdots n_k$.

This also gives a description of the elements of $F(\alpha_1, \alpha_2, \dots, \alpha_k)$. For simplicity, consider the case of the field $F(\alpha, \beta)$ where α and β are algebraic over F. Then the elements of this field are of the form

$$b_0 + b_1\beta + b_2\beta^2 + \cdots + b_{d-1}\beta^{d-1}$$

where $d = [F(\alpha)(\beta) : F(\alpha)]$ is the degree of β over $F(\alpha)$ (which may be strictly smaller than the degree of β over F), and where the coefficients b_0, $b_1, \dots, b_{d-1}$ are elements of $F(\alpha)$. The coefficients $b_i \in F(\alpha)$, $i = 0, \dots, d-1$, are of the form

$$a_{0i} + a_{1i}\alpha + a_{2i}\alpha^2 + \cdots + a_{n\text{-}1i}\alpha^{n-1}$$

where $n = [F(\alpha) : F]$ is the degree of α over F and the a_{ij} are elements of F. Hence the elements of $F(\alpha, \beta)$ are of the form

$$\sum_{\substack{i=0,1,\dots,n-1\\ j=0,1,\dots,d-1}} a_{ij}\alpha^i\beta^j \qquad\qquad a_{ij} \in F.$$

Since $[F(\alpha, \beta) : F] = [F(\alpha, \beta) : F(\alpha)][F(\alpha) : F] = dn$, the elements $\alpha^i\beta^j$ are in fact an F basis for $F(\alpha, \beta)$.

In practice the field $F(\alpha)$ generated by the algebraic α is obtained by adjoining the element α to F and then "closing" the resulting set with respect to addition and multiplication, which amounts to adjoining the powers $\alpha^2, \alpha^3, \dots$ of α and taking linear combinations (with coefficients from F) of these elements. The process terminates when a power of α is a linear combination of lower powers of α which amounts to knowing the minimal polynomial for α. The previous discussion shows a similar process gives the field $F(\alpha, \beta)$ generated by two elements, and by recursion, the field generated by any finite number of algebraic elements. This shows in particular that "closing" with respect to addition and multiplication also closes with respect to division for algebraic elements (cf. Example 5 following Corollary 5 above). If the elements are not algebraic, one must also "close" with respect to inverses. The difficulty in this procedure is determining the degrees of the *relative* extensions — for example the degree d for $F(\alpha, \beta)$ over $F(\alpha)$ above, for which one has only an a priori upper bound (the degree of β over F).

This is the analogue of "closing" a set of elements in a group G to determine the subgroup they generate.

Examples

(1) The extension $\mathbb{Q}(\sqrt[6]{2}, \sqrt{2})$ is simply the extension $\mathbb{Q}(\sqrt[6]{2})$ since $\sqrt{2}$ is already an element of this field. Put another way, the degree d of $\sqrt{2}$ over $\mathbb{Q}(\sqrt[6]{2})$ is 1, which is strictly smaller than the degree of $\sqrt{2}$ over $\mathbb{Q}$. We shall later have less obvious examples where this occurs.

(2) Consider the field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ generated over $\mathbb{Q}$ by $\sqrt{2}$ and $\sqrt{3}$. Since $\sqrt{3}$ is of degree 2 over $\mathbb{Q}$ the degree of the extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}(\sqrt{2})$ is at most 2 and is precisely 2 if and only if $x^2 - 3$ is irreducible over $\mathbb{Q}(\sqrt{2})$. Since this polynomial is of degree 2, it is reducible only if it has a root, i.e., if and only if $\sqrt{3} \in \mathbb{Q}(\sqrt{2})$. Suppose $\sqrt{3} = a + b\sqrt{2}$ with $a, b \in \mathbb{Q}$. Squaring this we obtain $3 = (a^2 + 2b^2) + 2ab\sqrt{2}$. If $ab \neq 0$, then we can solve this equation for $\sqrt{2}$ in terms of a and b which implies that $\sqrt{2}$ is rational, which it is not. If $b = 0$, then we would have that $\sqrt{3} = a$ is rational, a contradiction. Finally, if $a = 0$, we have $\sqrt{3} = b\sqrt{2}$ and multiplying both sides by $\sqrt{2}$ we see that $\sqrt{6}$ would be rational, again a contradiction. This shows $\sqrt{3} \notin \mathbb{Q}(\sqrt{2})$, proving
$$[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}] = 4.$$
Elements in this field (by "closing " 1, $\sqrt{2}$, $\sqrt{3}$) include 1, $\sqrt{2}$, $\sqrt{3}$, $\sqrt{6}$ and by the computations above, these form a basis for this field:
$$\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \{a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mid a, b, c, d \in \mathbb{Q}\}.$$

We can now characterize the finite extensions of a field F:

Theorem 17. The extension K/F is finite if and only if K is generated by a finite number of algebraic elements over F. More precisely, a field generated over F by a finite number of algebraic elements of degrees $n_1, n_2, \dots, n_k$ is algebraic of degree $\leq n_1 n_2 \cdots n_k$.

Proof: If K/F is finite of degree n, let $\alpha_1, \alpha_2, \dots, \alpha_n$ be a basis for K as a vector space over F. By Corollary 15, $[F(\alpha_i) : F]$ divides $[K : F] = n$ for $i = 1, 2, \dots, n$, so that Proposition 12 implies each α_i is algebraic over F. Since K is obviously generated over F by $\alpha_1, \alpha_2, \dots, \alpha_n$, we see that K is generated by a finite number of algebraic elements over F. The converse was proved above. The second statement of the theorem is immediate from Corollary 13 and the computation above.

The first example above shows that the inequality for the degree of the extension given in the theorem may be strict. We remark that information helpful in the determination of this degree can often be obtained by determining subfields and then applying Corollary 15.

Corollary 18. Suppose α and β are algebraic over F. Then $\alpha \pm \beta, \alpha\beta, \alpha/\beta$ (for $\beta \neq 0$), (in particular α^{-1} for $\alpha \neq 0$) are all algebraic.

Proof: All of these elements lie in the extension $F(\alpha, \beta)$, which is finite over F by the theorem, hence they are algebraic by Corollary 13.

Corollary 19. Let L/F be an arbitrary extension. Then the collection of elements of L that are algebraic over F form a subfield K of L.

Proof: This is immediate from the previous corollary.

Examples

(1) Consider the extension $\mathbb{C}/\mathbb{Q}$ and let $\overline{\mathbb{Q}}$ denote the subfield of all elements in $\mathbb{C}$ that are algebraic over $\mathbb{Q}$. In particular, the elements $\sqrt[n]{2}$ (the positive n^{th} roots of 2 in $\mathbb{R}$) are all elements of $\overline{\mathbb{Q}}$, so that $[\overline{\mathbb{Q}} : \mathbb{Q}] \geq n$ for all integers $n > 1$. Hence $\overline{\mathbb{Q}}$ is an *infinite* algebraic extension of $\mathbb{Q}$, called the field of *algebraic numbers*.

(2) Consider the field $\overline{\mathbb{Q}} \cap \mathbb{R}$, the subfield of $\mathbb{R}$ consisting of elements algebraic over $\mathbb{Q}$. The field $\mathbb{Q}$ is *countable*. The number of polynomials in $\mathbb{Q}[x]$ of any given degree n is therefore also countable (since such a polynomial is determined by specifying $n + 1$ coefficients from $\mathbb{Q}$). Since these polynomials have at most n roots in $\mathbb{R}$, the number of algebraic elements of $\mathbb{R}$ of degree n is countable. Finally, the collection of all algebraic elements in $\mathbb{R}$ is the countable union (indexed by n) of countable sets, hence is countable. Since $\mathbb{R}$ is uncountable, it follows that there exist (in fact many) elements of $\mathbb{R}$ which are not algebraic, i.e., are transcendental, over $\mathbb{Q}$. In particular the subfield $\overline{\mathbb{Q}} \cap \mathbb{R}$ of algebraic elements of $\mathbb{R}$ is a *proper* subfield of $\mathbb{R}$, so also $\overline{\mathbb{Q}}$ is a proper subfield of $\mathbb{C}$.

It is extremely difficult in general to prove that a given real number is not algebraic. For example, it is known (these are theorems) that $\pi = 3.14159...$ and $e = 2.71828...$ are transcendental elements of $\mathbb{R}$. Even the proofs that these elements are not *rational* are not too easy.

Theorem 20. If K is algebraic over F and L is algebraic over K, then L is algebraic over F.

Proof: Let α be any element of L. Then α is algebraic over K, so α satisfies some polynomial equation

$$a_n\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_1\alpha + a_0 = 0$$

where the coefficients $a_0, a_1, \dots, a_n$ are in K. Consider the field $F(\alpha, a_0, a_1, \dots, a_n)$ generated over F by α and the coefficients of this polynomial. Since K/F is algebraic, the elements $a_0, a_1, \dots, a_n$ are algebraic over F, so the extension $F(a_0, a_1, \dots, a_n)/F$ is finite by Theorem 17. By the equation above, we see that α generates an extension of this field of degree at most n, since its minimal polynomial over this field is a divisor of the polynomial above. Therefore

$$[F(\alpha, a_0, a_1, \dots, a_n) : F] = [F(\alpha, a_0, \dots, a_n) : F(a_0, \dots, a_n)][F(a_0, \dots, a_n) : F]$$

is also finite and $F(\alpha, a_0, a_1, \dots, a_n)/F$ is an algebraic extension. In particular the element α is algebraic over F, which proves that L is algebraic over F.

The subfield $F(\alpha_1, \alpha_2, \dots, \alpha_k)$ generated by a finite set of elements $\alpha_1, \alpha_2, \dots, \alpha_k$ of a field K contains each of the fields $F(\alpha_i)$, $i = 1, 2, \dots, k$. By the definitions, it is also the smallest subfield of K containing these fields.

Definition. Let K_1 and K_2 be two subfields of a field K. Then the *composite field* of K_1 and K_2, denoted K_1K_2, is the smallest subfield of K containing both K_1 and K_2. Similarly, the composite of any collection of subfields of K is the smallest subfield containing all the subfields.

Note that the composite K_1K_2 can also be described as the intersection of all the subfields of K containing both K_1 and K_2 and similarly for the composite of more than two fields, analogous to the subgroup generated by a subset of a group (cf. Section 2.4).

Example

The composite of the two fields $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt[3]{2})$ is the field $\mathbb{Q}(\sqrt[6]{2})$. This is because this field contains both of these subfields ($(\sqrt[6]{2})^3 = \sqrt{2}$ and $(\sqrt[6]{2})^2 = \sqrt[3]{2}$) and conversely, any field containing both $\sqrt{2}$ and $\sqrt[3]{2}$ contains their quotient, which is $\sqrt[6]{2}$.

Suppose now that K_1 and K_2 are finite extensions of F in K. Let $\alpha_1, \alpha_2, \dots, \alpha_n$ be an F-basis for K_1 and let $\beta_1, \beta_2, \dots, \beta_m$ be an F-basis for K_2 (so that $[K_1 : F] = n$ and $[K_2 : F] = m$). Then it is clear that these give generators for the composite K_1K_2 over F:

$$K_1K_2 = F(\alpha_1, \alpha_2, \dots, \alpha_n, \beta_1, \beta_2, \dots, \beta_m).$$

Since $\alpha_1, \alpha_2, \dots, \alpha_n$ is an F-basis for K_1 any power ${\alpha_i}^k$ of one of the α's is a *linear* combination with coefficients in F of the α's and a similar statement holds for the β's. It follows that the collection of linear combinations

$$\sum_{\substack{i=1,2,\dots,n\\ j=1,2,\dots,m}} a_{ij}\alpha_i\beta_j$$

with coefficients in F is *closed* under multiplication and addition since in a product of two such elements any higher powers of the α's and β's can be replaced by linear expressions. Hence, the elements $\alpha_i\beta_j$ for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$ *span* the composite extension K_1K_2 over F. In particular, $[K_1K_2 : F] \le mn$. We summarize this as:

Proposition 21. Let K_1 and K_2 be two finite extensions of a field F contained in K. Then

$$[K_1K_2 : F] \leq [K_1 : F][K_2 : F]$$

with equality if and only if an F-basis for one of the fields remains linearly independent over the other field. If $\alpha_1, \alpha_2, \dots, \alpha_n$ and $\beta_1, \beta_2, \dots, \beta_m$ are bases for K_1 and K_2 over F, respectively, then the elements $\alpha_i\beta_j$ for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$ span K_1K_2 over F.

Proof: From $K_1K_2 = F(\alpha_1, \alpha_2, \dots, \alpha_n, \beta_1, \beta_2, \dots, \beta_m) = K_1(\beta_1, \beta_2, \dots, \beta_m)$, we see as above that $\beta_1, \beta_2, \dots, \beta_m$ span K_1K_2 over K_1. Hence $[K_1K_2 : K_1] \leq m = [K_2 : F]$ with equality if and only if these elements are linearly independent over K_1. Since $[K_1K_2 : F] = [K_1K_2 : K_1][K_1 : F]$ this proves the proposition.

By the proposition (and its proof), we have the following diagram:

$$\begin{tikzcd}
& K_1K_2 & \\
K_1 \arrow[ur, dash, "\leq m"] & & K_2 \arrow[ul, dash, "\leq n"'] \\
& F \arrow[ul, dash, "n"] \arrow[ur, dash, "m"'] &
\end{tikzcd}$$

We shall have examples shortly where the inequality in the proposition is strict. One useful situation where one can be certain of equality is the following:

Corollary 22. Suppose that $[K_1 : F] = n$, $[K_2 : F] = m$ in Proposition 21, where n and m are relatively prime: $(n, m) = 1$. Then $[K_1K_2 : F] = [K_1 : F][K_2 : F] = nm$.

Proof: In general the extension degree $[K_1K_2 : F]$ is divisible by both n and m since K_1 and K_2 are subfields of K_1K_2, hence is divisible by their least common multiple. In this case, since $(n, m) = 1$, this means $[K_1K_2 : F]$ is divisible by nm, which together with the inequality $[K_1K_2 : F] \leq nm$ of the proposition proves the corollary.

Example

The composite of the two fields $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt[3]{2})$ is of degree 6 over $\mathbb{Q}$, which we determined earlier by actually computing the composite $\mathbb{Q}(\sqrt[6]{2})$.

EXERCISES

1. Let $\mathbb{F}$ be a finite field of characteristic p. Prove that $|\mathbb{F}| = p^n$ for some positive integer n.
2. Let $g(x) = x^2 + x - 1$ and let $h(x) = x^3 - x + 1$. Obtain fields of 4, 8, 9 and 27 elements by adjoining a root of $f(x)$ to the field F where $f(x) = g(x)$ or $h(x)$ and $F = \mathbb{F}_2$ or $\mathbb{F}_3$. Write down the multiplication tables for the fields with 4 and 9 elements and show that the nonzero elements form a cyclic group.
3. Determine the minimal polynomial over $\mathbb{Q}$ for the element $1 + i$.

4. Determine the degree over $\mathbb{Q}$ of $2+\sqrt{3}$ and of $1+\sqrt[3]{2}+\sqrt[3]{4}$.

5. Let $F=\mathbb{Q}(i)$. Prove that x^3-2 and x^3-3 are irreducible over F.

6. Prove directly from the definitions that the field $F(\alpha_1,\alpha_2,\dots,\alpha_n)$ is the composite of the fields $F(\alpha_1), F(\alpha_2),\dots,F(\alpha_n)$.

7. Prove that $\mathbb{Q}(\sqrt{2}+\sqrt{3})=\mathbb{Q}(\sqrt{2},\sqrt{3})$ [one inclusion is obvious, for the other consider $(\sqrt{2}+\sqrt{3})^2$, etc.]. Conclude that $[\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}]=4$. Find an irreducible polynomial satisfied by $\sqrt{2}+\sqrt{3}$.

8. Let F be a field of characteristic $\neq 2$. Let D_1 and D_2 be elements of F, neither of which is a square in F. Prove that $F(\sqrt{D_1},\sqrt{D_2})$ is of degree 4 over F if D_1D_2 is not a square in F and is of degree 2 over F otherwise. When $F(\sqrt{D_1},\sqrt{D_2})$ is of degree 4 over F the field is called a *biquadratic extension of F*.

9. Let F be a field of characteristic $\neq 2$. Let a, b be elements of the field F with b not a square in F. Prove that a necessary and sufficient condition for $\sqrt{a+\sqrt{b}}=\sqrt{m}+\sqrt{n}$ for some m and n in F is that a^2-b is a square in F. Use this to determine when the field $\mathbb{Q}(\sqrt{a+\sqrt{b}})$ $(a,b\in\mathbb{Q})$ is biquadratic over $\mathbb{Q}$.

10. Determine the degree of the extension $\mathbb{Q}(\sqrt{3+2\sqrt{2}})$ over $\mathbb{Q}$.

11. **(a)** Let $\sqrt{3+4i}$ denote the square root of the complex number $3+4i$ that lies in the first quadrant and let $\sqrt{3-4i}$ denote the square root of $3-4i$ that lies in the fourth quadrant. Prove that $[\mathbb{Q}(\sqrt{3+4i}+\sqrt{3-4i}):\mathbb{Q}]=1$.

(b) Determine the degree of the extension $\mathbb{Q}(\sqrt{1+\sqrt{-3}}+\sqrt{1-\sqrt{-3}})$ over $\mathbb{Q}$.

12. Suppose the degree of the extension K/F is a prime p. Show that any subfield E of K containing F is either K or F.

13. Suppose $F=\mathbb{Q}(\alpha_1,\alpha_2,\dots,\alpha_n)$ where $\alpha_i^2\in\mathbb{Q}$ for $i=1,2,\dots,n$. Prove that $\sqrt[3]{2}\notin F$.

14. Prove that if $[F(\alpha):F]$ is odd then $F(\alpha)=F(\alpha^2)$.

15. A field F is said to be *formally real* if -1 is not expressible as a sum of squares in F. Let F be a formally real field, let $f(x)\in F[x]$ be an irreducible polynomial of odd degree and let α be a root of $f(x)$. Prove that $F(\alpha)$ is also formally real. [Pick α a counterexample of minimal degree. Show that $-1+f(x)g(x)=(p_1(x))^2+\cdots+(p_m(x))^2$ for some $p_i(x), g(x)\in F[x]$ where $g(x)$ has odd degree $<\deg f$. Show that some root β of g has odd degree over F and $F(\beta)$ is not formally real, violating the minimality of α.]

16. Let K/F be an algebraic extension and let R be a *ring* contained in K and containing F. Show that R is a subfield of K containing F.

17. Let $f(x)$ be an irreducible polynomial of degree n over a field F. Let $g(x)$ be any polynomial in $F[x]$. Prove that every irreducible factor of the composite polynomial $f(g(x))$ has degree divisible by n.

18. Let k be a field and let $k(x)$ be the field of rational functions in x with coefficients from k. Let $t\in k(x)$ be the rational function $\dfrac{P(x)}{Q(x)}$ with relatively prime polynomials $P(x), Q(x)\in k[x]$, with $Q(x)\neq 0$. Then $k(x)$ is an extension of $k(t)$ and to compute its degree it is necessary to compute the minimal polynomial with coefficients in $k(t)$ satisfied by x.

(a) Show that the polynomial $P(X)-tQ(X)$ in the variable X and coefficients in $k(t)$ is irreducible over $k(t)$ and has x as a root. [By Gauss' Lemma this polynomial is irreducible in $(k(t))[X]$ if and only if it is irreducible in $(k[t])[X]$. Then note that $(k[t])[X]=(k[X])[t]$.]

(b) Show that the degree of $P(X) - tQ(X)$ as a polynomial in X with coefficients in $k(t)$ is the maximum of the degrees of $P(x)$ and $Q(x)$.

(c) Show that $[k(x) : k(t)] = [k(x) : k(\frac{P(x)}{Q(x)})] = \max(\deg P(x), \deg Q(x))$.

19. Let K be an extension of F of degree n.

(a) For any $\alpha \in K$ prove that α acting by left multiplication on K is an F-linear transformation of K.

(b) Prove that K is isomorphic to a subfield of the ring of $n \times n$ matrices over F, so the ring of $n \times n$ matrices over F contains an isomorphic copy of *every* extension of F of degree $\leq n$.

20. Show that if the matrix of the linear transformation "multiplication by α" considered in the previous exercise is A then α is a root of the characteristic polynomial for A. This gives an effective procedure for determining an equation of degree n satisfied by an element α in an extension of F of degree n. Use this procedure to obtain the monic polynomial of degree 3 satisfied by $\sqrt[3]{2}$ and by $1 + \sqrt[3]{2} + \sqrt[3]{4}$.

21. Let $K = \mathbb{Q}(\sqrt{D})$ for some squarefree integer D. Let $\alpha = a + b\sqrt{D}$ be an element of K. Use the basis $1, \sqrt{D}$ for K as a vector space over $\mathbb{Q}$ and show that the matrix of the linear transformation "multiplication by α" on K considered in the previous exercises has the matrix $\begin{pmatrix} a & bD \\ b & a \end{pmatrix}$. Prove directly that the map $a + b\sqrt{D} \mapsto \begin{pmatrix} a & bD \\ b & a \end{pmatrix}$ is an isomorphism of the field K with a subfield of the ring of 2×2 matrices with coefficients in $\mathbb{Q}$.

22. Let K_1 and K_2 be two finite extensions of a field F contained in the field K. Prove that the F-algebra $K_1 \otimes_F K_2$ is a field if and only if $[K_1K_2 : F] = [K_1 : F][K_2 : F]$.

13.3 CLASSICAL STRAIGHTEDGE AND COMPASS CONSTRUCTIONS

As a simple application of the results we have obtained on algebraic extensions, and in particular on the multiplicativity of extension degrees, we can answer (in the negative) the following geometric problems posed by the Greeks:

I. *(Doubling the Cube)* Is it possible using only straightedge and compass to construct a cube with precisely twice the volume of a given cube?

II. *(Trisecting an Angle)* Is it possible using only straightedge and compass to trisect any given angle θ?

III. *(Squaring the Circle)* Is it possible using only straightedge and compass to construct a square whose area is precisely the area of a given circle?

To answer these questions we must translate the construction of lengths by compass and straightedge into algebraic terms. Let 1 denote a fixed given unit distance. Then any distance is determined by its length $a \in \mathbb{R}$, which allows us to view geometric distances as elements of the real numbers $\mathbb{R}$. Using the given unit distance 1 to define the scale on the axes, we can then construct the usual Cartesian plane $\mathbb{R}^2$ and view all of our constructions as occurring in $\mathbb{R}^2$. A point $(x, y) \in \mathbb{R}^2$ is then constructible starting with the given distance 1 if and only if its coordinates x and y are constructible elements of $\mathbb{R}$. The problems above then amount to determining whether particular lengths in $\mathbb{R}$ can be obtained by compass and straightedge constructions from a fixed unit distance. The collection of such real numbers together with their negatives will be called the *constructible* elements of $\mathbb{R}$, and we shall not distinguish between the lengths that are constructible and the real numbers that are constructible.

Each straightedge and compass construction consists of a series of operations of the following four types: (1) connecting two given points by a straight line, (2) finding a point of intersection of two straight lines, (3) drawing a circle with given radius and center, and (4) finding the point(s) of intersection of a straight line and a circle or the intersection of two circles.

It is an elementary fact from geometry that if two lengths a and b are given one may construct using straightedge and compass the lengths $a \pm b$, ab and a/b (the first two are clear and the latter two are given by the construction of parallel lines (Figure 1)).

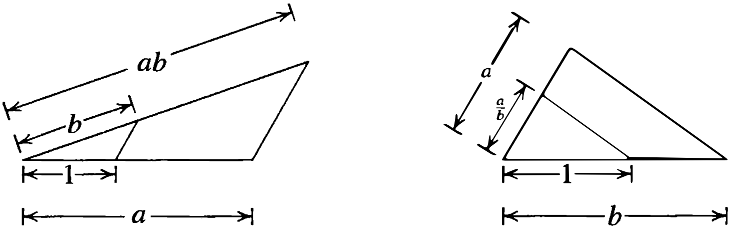


Fig. 1

It is also an elementary geometry construction to construct $\sqrt{a}$ if a is given: construct the circle with diameter $1 + a$ and erect the perpendicular to the diameter as indicated in Figure 2. Then $\sqrt{a}$ is the length of this perpendicular.

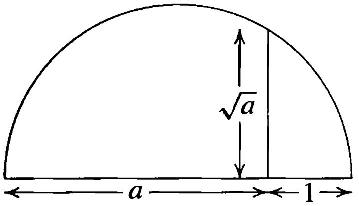


Fig. 2

It follows that straightedge and compass constructions give all the algebraic operations of addition, subtraction, multiplication and division (by nonzero elements) in the reals so the collection of constructible elements is a *subfield* of $\mathbb{R}$. One can also take square roots of constructible elements. We shall now see that these are essentially the only operations possible.

From the given length 1 it is possible to construct by these operations all the rational numbers $\mathbb{Q}$. Hence we may construct all of the points $(x, y) \in \mathbb{R}^2$ whose coordinates are rational. We may construct additional elements of $\mathbb{R}$ by taking square roots, so the collection of elements constructible from 1 of $\mathbb{R}$ form a field strictly larger than $\mathbb{Q}$.

The usual formula ("two point form") for the straight line connecting two points with coordinates in some field F gives an equation for the line of the form $ax+by-c = 0$ with $a, b, c \in F$. Solving two such equations simultaneously to determine the point of intersection of two such lines gives solutions also in F. It follows that if the coordinates of two points lie in the field F then straightedge constructions alone will not produce additional points whose coordinates are not also in F.

A compass construction (type (3) or (4) above) defines points obtained by the intersection of a circle with either a straight line or another circle. A circle with center (h, k) and radius r has equation

$$(x-h)^2+(y-k)^2=r^2$$

so when we consider the effect of compass constructions on elements of a field F we are considering simultaneous solutions of such an equation with a linear equation $ax+by-c=0$ where $a, b, c, h, k, r \in F$, or the simultaneous solutions of two quadratic equations.

In the case of a linear equation and the equation for the circle, solving for y, say, in the linear equation and substituting gives a *quadratic* equation for x (and y is given linearly in terms of x). Hence the coordinates of the point of intersection are at worst in a *quadratic extension* of F.

In the case of the intersection of two circles, say

$$\begin{aligned} &(x-h)^2+(y-k)^2=r^2 \\ \text{and} \qquad &(x-h')^2+(y-k')^2=r'^2, \end{aligned}$$

subtraction of the second equation from the first shows that we have the same intersection by considering the two equations

$$\begin{aligned} &(x-h)^2+(y-k)^2=r^2 \\ \text{and} \qquad &2(h'-h)x+2(k'-k)y=r^2-h^2-k^2-r'^2+h'^2+k'^2 \end{aligned}$$

which is the intersection of a circle and a straight line (the straight line connecting the two points of intersection, in fact) of the type just considered.

It follows that if a collection of constructible elements is given, then one can construct all the elements in the subfield F of $\mathbb{R}$ generated by these elements and that any straightedge and compass operation on elements of F produces elements in at worst a *quadratic* extension of F. Since quadratic extensions have degree 2 and extension degrees are multiplicative, it follows that if $\alpha \in \mathbb{R}$ is obtained from elements in a field F by a (finite) series of straightedge and compass operations then α is an element of an extension K of F of degree a power of 2: $[K : F] = 2^m$ for some m. Since $[F(\alpha) : F]$ divides this extension degree, it must also be a power of 2.

Proposition 23. If the element $\alpha \in \mathbb{R}$ is obtained from a field $F \subset \mathbb{R}$ by a series of compass and straightedge constructions then $[F(\alpha) : F] = 2^k$ for some integer $k \geq 0$.

Theorem 24. None of the classical Greek problems: (I) Doubling the Cube, (II) Trisecting an Angle, and (III) Squaring the Circle, is possible.

Proof: (I) Doubling the cube amounts to constructing $\sqrt[3]{2}$ in the reals starting with the unit 1. Since $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$ is not a power of 2, this is impossible.

(II) If an angle θ can be constructed, then determining the point at distance 1 from the origin and angle θ from the positive x axis in $\mathbb{R}^2$ shows that $\cos\theta$ (the x-coordinate of this point) can be constructed (so then $\sin\theta$ can also be constructed). Conversely if $\cos\theta$, then $\sin\theta$, can be constructed, the point with those coordinates gives the angle θ.

The problem of trisecting the angle θ is then equivalent to the problem: given $\cos\theta$ construct $\cos\theta/3$.

To see that this is not always possible (it is certainly occasionally possible, for example for $\theta = 180^\circ$), consider $\theta = 60^\circ$. Then $\cos\theta = \frac{1}{2}$. By the triple angle formula for cosines:

$$\cos\theta = 4\cos^3\theta/3 - 3\cos\theta/3,$$

substituting $\theta = 60^\circ$, we see that $\beta = \cos 20^\circ$ satisfies the equation

$$4\beta^3 - 3\beta - 1/2 = 0$$

or $8(\beta)^3 - 6\beta - 1 = 0$. This can be written $(2\beta)^3 - 3(2\beta) - 1 = 0$. Let $\alpha = 2\beta$. Then α is a real number between 0 and 2 satisfying the equation

$$\alpha^3 - 3\alpha - 1 = 0.$$

But we considered this equation in the last section and determined $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$, and as before we see that α is not constructible.

(III) Squaring the circle is equivalent to determining whether the real number $\pi = 3.14159\ldots$ is constructible. As mentioned previously, it is a difficult problem even to prove that this number is not rational. It is in fact transcendental (which we shall assume without proof), so that $[\mathbb{Q}(\pi) : \mathbb{Q}]$ is not even finite, much less a power of 2, showing the impossibility of squaring the circle by straightedge and compass.

Remark: The proof above shows that $\cos 20^\circ$ and $\sin 20^\circ$ cannot be constructed. The question arises as to which integer angles (measured in degrees) are constructible? The angles 1° and 2° are not constructible, since otherwise the addition formulae for sines and cosines would give the constructibility for 20°. On the other hand, elementary geometric constructions (of the regular 5-gon for an angle of 72° and the equilateral triangle for an angle of 60°) together with the addition formulae and the half-angle formulae show that $\cos 3^\circ$ and $\sin 3^\circ$ are constructible. It follows from this that the trigonometric functions of an integer degree angle are constructible precisely when the angle is a multiple of 3°. Explicitly,

$$\cos 3^\circ = \frac{1}{8}(\sqrt{3}+1)\sqrt{5+\sqrt{5}} + \frac{1}{16}(\sqrt{6}-\sqrt{2})(\sqrt{5}-1)$$
$$\sin 3^\circ = \frac{1}{16}(\sqrt{6}+\sqrt{2})(\sqrt{5}-1) - \frac{1}{8}(\sqrt{3}-1)\sqrt{5+\sqrt{5}},$$

showing that these are obtained from $\mathbb{Q}$ by successive extractions of square roots and field operations.

After discussing the cyclotomic fields in Section 14.5 we shall consider another classical geometric question: "which regular n-gons can be constructed by straightedge and compass?" (cf. Proposition 14.29).

We have been careful here to consider constructions using a *straightedge* rather than a *ruler*, the distinction being that a ruler has marks on it. If one uses a ruler, it is possible to construct many additional algebraic elements. For example, suppose θ is a given angle and the unit distance 1 is marked on the ruler. Draw a circle of radius 1 with central angle θ as shown in Figure 3 and then slide the ruler until the distance between points A and B on the circle is 1. Then some elementary geometry shows that (cf. the exercises) the angle α indicated is $\theta/3$, i.e., this construction (due to Archimedes) trisects θ. In particular, the second classical problem in Theorem 24 (Trisecting an Angle) can be solved with *ruler* and compass.

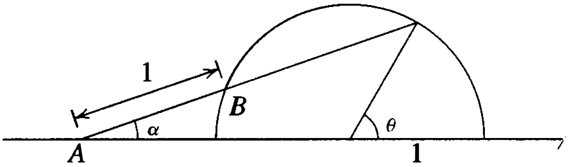


Fig. 3

The first of the classical problems in Theorem 24 (Duplication of the Cube), which amounts to the construction of $\sqrt[3]{2}$, can also be solved with ruler and compass. The following gives a construction for $k^{1/3}$ for any given positive real k which is less than 1. This construction was shown to us by J.H. Conway.

Drawing a circle of radius 1 and using the point $A = (k, 0)$ as center, construct the point $B = (0, \sqrt{1-k^2})$. Dividing this distance by 3, construct the point $(0, -\frac{1}{3}\sqrt{1-k^2})$ and draw the line connecting this point with A. Slide the ruler with marked unit length 1 so that it passes through the point B and so that the distance from the intersection point C to the intersection point D with the x-axis is of length 1, as indicated in Figure 4.

Then the distance between A and D is $2k^{1/3}$ and the distance between B and C is $2k^{2/3}$ (cf. the exercises).

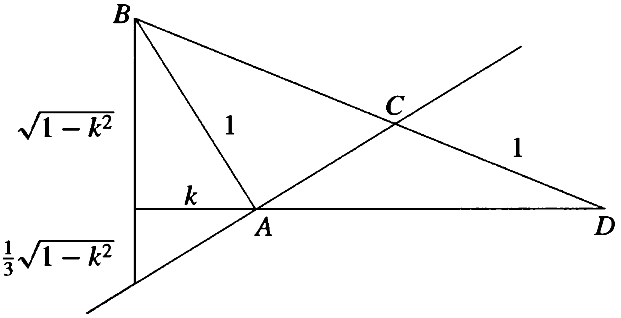


Fig. 4

EXERCISES

1. Prove that it is impossible to construct the regular 9-gon.
2. Prove that Archimedes' construction actually trisects the angle θ. [Note the isosceles triangles in Figure 5 to prove that $\beta = \gamma = 2\alpha$.]

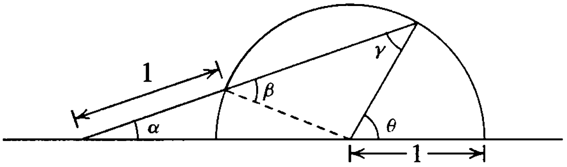


Fig. 5

3. Prove that Conway's construction indicated in the text actually constructs $2k^{1/3}$ and $2k^{2/3}$. [One method: let (x, y) be the coordinates of the point C, a the distance from B to C and b the distance from A to D; use similar triangles to prove (a) $\frac{y}{1} = \frac{\sqrt{1-k^2}}{1+a}$, (b) $\frac{x}{a} = \frac{b+k}{1+a}$, (c) $\frac{y}{x-k} = \frac{\sqrt{1-k^2}}{3k}$, and also show that (d) $(1-k^2)+(b+k)^2 = (1+a)^2$; solve these equations for a and b.]

4. The construction of the regular 7-gon amounts to the constructibility of $\cos(2\pi/7)$. We shall see later (Section 14.5 and Exercise 2 of Section 14.7) that $\alpha = 2\cos(2\pi/7)$ satisfies the equation $x^3 + x^2 - 2x - 1 = 0$. Use this to prove that the regular 7-gon is not constructible by straightedge and compass.

5. Use the fact that $\alpha = 2\cos(2\pi/5)$ satisfies the equation $x^2 + x - 1 = 0$ to conclude that the regular 5-gon is constructible by straightedge and compass.

13.4 SPLITTING FIELDS AND ALGEBRAIC CLOSURES

Let F be a field.

If $f(x)$ is any polynomial in $F[x]$ then we have seen in Section 2 that there exists a field K which can (by identifying F with an isomorphic copy of F) be considered an extension of F in which $f(x)$ has a root α. This is equivalent to the statement that $f(x)$ has a linear factor $x - \alpha$ in $K[x]$ (this is Proposition 9 of Chapter 9).

Definition. The extension field K of F is called a *splitting field* for the polynomial $f(x) \in F[x]$ if $f(x)$ factors completely into linear factors (or *splits completely*) in $K[x]$ and $f(x)$ does not factor completely into linear factors over any proper subfield of K containing F.

If $f(x)$ is of degree n, then $f(x)$ has at most n roots in F (Proposition 17 of Chapter 9) and has precisely n roots (counting multiplicities) in F if and only if $f(x)$ splits completely in $F[x]$.

Theorem 25. For any field F, if $f(x) \in F[x]$ then there exists an extension K of F which is a splitting field for $f(x)$.

Proof: We first show that there is an extension E of F over which $f(x)$ splits completely into linear factors by induction on the degree n of $f(x)$. If $n = 1$, then take $E = F$. Suppose now that $n > 1$. If the irreducible factors of $f(x)$ over F are all of degree 1, then F is the splitting field for $f(x)$ and we may take $E = F$. Otherwise, at least one of the irreducible factors, say $p(x)$ of $f(x)$ in $F[x]$ is of degree at least 2. By Theorem 3 there is an extension E_1 of F containing a root α of $p(x)$. Over E_1 the polynomial $f(x)$ has the linear factor $x - \alpha$. The degree of the remaining factor $f_1(x)$ of $f(x)$ is $n - 1$, so by induction there is an extension E of E_1 containing all the roots of $f_1(x)$. Since $\alpha \in E$, E is an extension of F containing all the roots of $f(x)$. Now let K be the intersection of all the subfields of E containing F which also contain all the roots of $f(x)$. Then K is a field which is a splitting field for $f(x)$.

We shall see shortly that any two splitting fields for $f(x)$ are isomorphic (which extends Theorem 8), so (by abuse) we frequently refer to *the* splitting field of a polynomial.

Definition. If K is an algebraic extension of F which is the splitting field over F for a collection of polynomials $f(x) \in F[x]$ then K is called a *normal* extension of F.

We shall generally use the term "splitting field" rather than "normal extension" (cf. also Section 14.9).

Examples

(1) The splitting field for $x^2 - 2$ over $\mathbb{Q}$ is just $\mathbb{Q}(\sqrt{2})$, since the two roots are $\pm\sqrt{2}$ and $-\sqrt{2} \in \mathbb{Q}(\sqrt{2})$.

(2) The splitting field for $(x^2 - 2)(x^2 - 3)$ is the field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ generated over $\mathbb{Q}$ by $\sqrt{2}$ and $\sqrt{3}$ since the roots of the polynomial are $\pm\sqrt{2}, \pm\sqrt{3}$. We have already seen that this is an extension of degree 4 over $\mathbb{Q}$ and we have the following diagram of known subfields:

$$\begin{tikzcd}
 & \mathbb{Q}(\sqrt{2}, \sqrt{3}) \arrow[dl, dash, "2"'] \arrow[d, dash, "2"] \arrow[dr, dash, "2"] & \\
\mathbb{Q}(\sqrt{2}) \arrow[dr, dash, "2"'] & \mathbb{Q}(\sqrt{6}) \arrow[d, dash, "2"] & \mathbb{Q}(\sqrt{3}) \arrow[dl, dash, "2"] \\
 & \mathbb{Q} &
\end{tikzcd}$$

(3) The splitting field of $x^3 - 2$ over $\mathbb{Q}$ is not just $\mathbb{Q}(\sqrt[3]{2})$ since as previously noted the three roots of this polynomial in $\mathbb{C}$ are

$$\sqrt[3]{2}, \quad \sqrt[3]{2}\left(\frac{-1 + i\sqrt{3}}{2}\right), \quad \sqrt[3]{2}\left(\frac{-1 - i\sqrt{3}}{2}\right)$$

and the latter two roots are not elements of $\mathbb{Q}(\sqrt[3]{2})$, since the elements of this field are of the form $a + b\sqrt[3]{2} + c\sqrt[3]{4}$ with rational a, b, c and all such numbers are real.

The splitting field K of this polynomial is obtained by adjoining all three of these roots to $\mathbb{Q}$. Note that since K contains the first two roots above, then it contains their quotient $\dfrac{-1 + \sqrt{-3}}{2}$ hence K contains the element $\sqrt{-3}$. On the other hand, any field containing $\sqrt[3]{2}$ and $\sqrt{-3}$ contains all three of the roots above. It follows that

$$K = \mathbb{Q}(\sqrt[3]{2}, \sqrt{-3})$$

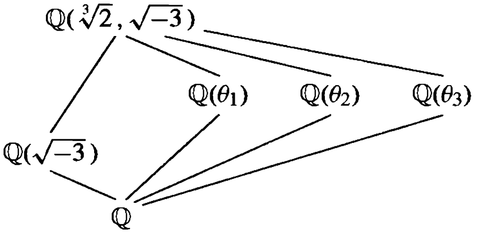
is the splitting field of $x^3 - 2$ over $\mathbb{Q}$. Since $\sqrt{-3}$ satisfies the equation $x^2 + 3 = 0$, the degree of this extension over $\mathbb{Q}(\sqrt[3]{2})$ is at most 2, hence must be 2 since we observed above that $\mathbb{Q}(\sqrt[3]{2})$ is not the splitting field. It follows that

$$[\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3}) : \mathbb{Q}] = 6.$$

Note that we could have proceeded slightly differently at the end by noting that $\mathbb{Q}(\sqrt{-3})$ is a subfield of K, so that the index $[\mathbb{Q}(\sqrt{-3}) : \mathbb{Q}] = 2$ divides $[K : \mathbb{Q}]$.

Since this extension degree is also divisible by 3 (because $\mathbb{Q}(\sqrt[3]{2}) \subset K$), the degree is divisible by 6, hence must be 6.

This gives us the diagram of known subfields:



where

$$\theta_1 = \sqrt[3]{2}, \quad \theta_2 = \sqrt[3]{2}\left(\frac{-1+i\sqrt{3}}{2}\right), \quad \theta_3 = \sqrt[3]{2}\left(\frac{-1-i\sqrt{3}}{2}\right).$$

(4) One must be careful in computing splitting fields. The splitting field for the polynomial x^4+4 over $\mathbb{Q}$ is smaller than one might at first suspect. In fact this polynomial factors over $\mathbb{Q}$:

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

where these two factors are irreducible (Eisenstein again). Solving for the roots of the two factors by the quadratic formula, we find the four roots

$$\pm 1 \pm i$$

so that the splitting field of this polynomial is just the field $\mathbb{Q}(i)$, an extension of degree 2 of $\mathbb{Q}$.

In general, if $f(x) \in F[x]$ is a polynomial of degree n, then adjoining one root of $f(x)$ to F generates an extension F_1 of degree at most n (and equal to n if and only if $f(x)$ is irreducible). Over F_1 the polynomial $f(x)$ now has at least one linear factor, so that any other root of $f(x)$ satisfies an equation of degree at most $n-1$ over F_1. Adjoining such a root to F_1 we therefore obtain an extension of degree at most $n-1$ of F_1, etc. Using the multiplicativity of extension degrees, this proves

Proposition 26. A splitting field of a polynomial of degree n over F is of degree at most $n!$ over F.

As the examples above show, the degree of a splitting field may be smaller than $n!$. It will be proved later using Galois Theory that a "general" polynomial of degree n (in a well defined sense) over $\mathbb{Q}$ has a splitting field of degree $n!$, so this may be viewed as the "generic" situation (although most of the interesting examples we shall consider have splitting fields of smaller degree).

Example: (Splitting Field of $x^n - 1$: Cyclotomic Fields)

Consider the splitting field of the polynomial $x^n - 1$ over $\mathbb{Q}$. The roots of this polynomial are called the n^{th} *roots of unity*.

Recall that every nonzero complex number $a + bi \in \mathbb{C}$ can be written uniquely in the form

$$re^{i\theta} = r(\cos\theta + i\sin\theta) \qquad r > 0, \quad 0 \leq \theta < 2\pi$$

which is simply representing the point $a + bi$ in the complex plane in terms of polar coordinates: r is the distance of (a, b) from the origin and θ is the angle made with the real positive axis.

Over $\mathbb{C}$ there are n distinct solutions of the equation $x^n = 1$, namely the elements

$$e^{2\pi ki/n} = \cos\left(\frac{2\pi k}{n}\right) + i\sin\left(\frac{2\pi k}{n}\right)$$

for $k = 0, 1, \ldots, n-1$. These points are given geometrically by n equally spaced points starting with the point (1,0) (corresponding to $k = 0$) on a circle of radius 1 in the complex plane (see Figure 6). The fact that these are all n^{th} roots of unity is immediate, since

$$(e^{2\pi ki/n})^n = e^{(2\pi ki/n)n} = e^{2\pi ki} = 1.$$

It follows that $\mathbb{C}$ contains a splitting field for $x^n - 1$ and we shall frequently view the splitting field for $x^n - 1$ over $\mathbb{Q}$ as the field generated over $\mathbb{Q}$ in $\mathbb{C}$ by the numbers above.

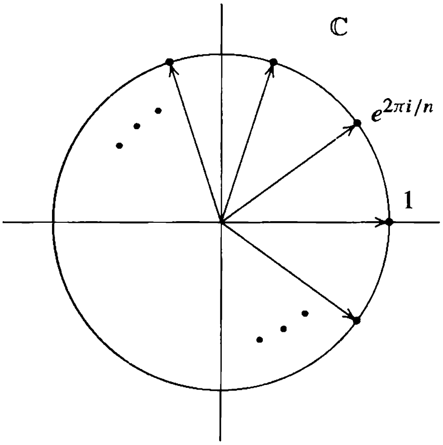


Fig. 6

In *any* abstract splitting field $K/\mathbb{Q}$ for $x^n - 1$ the collection of n^{th} roots of unity form a *group* under multiplication since if $\alpha^n = 1$ and $\beta^n = 1$ then $(\alpha\beta)^n = 1$, so this subset of $K^\times$ is closed under multiplication. It follows that this is a *cyclic* group (Proposition 18 of Chapter 9); we shall see that there are n distinct roots in K so it has order n.

Definition. A generator of the cyclic group of all n^{th} roots of unity is called a *primitive* n^{th} root of unity.

Let ζ_n denote a primitive n^{th} root of unity. The other *primitive* n^{th} roots of unity are then the elements $\zeta_n{}^a$ where $1 \leq a < n$ is an integer relatively prime to n, since these are the other generators for a cyclic group of order n. In particular there are precisely $\varphi(n)$ primitive n^{th} roots of unity, where $\varphi(n)$ denotes the Euler φ-function.

Over $\mathbb{C}$ we can see all of this directly by letting

$$\zeta_n = e^{2\pi i/n}$$

(the first n^{th} root of unity counterclockwise from 1). Then all the other roots of unity are powers of ζ_n:

$$e^{2\pi k i/n} = \zeta_n^k$$

so that ζ_n is one possible generator for the multiplicative group of n^{th} roots of unity. When we view the roots of unity in $\mathbb{C}$ we shall usually use ζ_n to denote this choice of a primitive n^{th} root of unity. The primitive roots of unity in $\mathbb{C}$ for some small values of n are

$$\begin{aligned}
\zeta_1 &= 1\\
\zeta_2 &= -1\\
\zeta_3 &= \frac{-1+i\sqrt{3}}{2}\\
\zeta_4 &= i\\
\zeta_5 &= \frac{\sqrt{5}-1}{4} + i\left(\frac{\sqrt{10+2\sqrt{5}}}{4}\right)\\
\zeta_6 &= \frac{1+i\sqrt{3}}{2}\\
\zeta_8 &= \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}
\end{aligned}$$

(these formulas follow from the elementary geometry of n-gons and in any case can be verified directly by raising them to the appropriate power).

The splitting field of $x^n - 1$ over $\mathbb{Q}$ is the field $\mathbb{Q}(\zeta_n)$ and this field is given a name:

Definition. The field $\mathbb{Q}(\zeta_n)$ is called the *cyclotomic field of n^{th} roots of unity.*

Determining the degree of this extension requires some analysis of the minimal polynomial of ζ_n over $\mathbb{Q}$ and will be postponed until later (Section 6). One important special case which we have in fact already considered is when $n = p$ is a *prime*. In this case, we have the factorization

$$x^p - 1 = (x-1)(x^{p-1} + x^{p-2} + \cdots + x + 1)$$

and since $\zeta_p \neq 1$ it follows that ζ_p is a root of the polynomial

$$\Phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + x + 1$$

which we showed was irreducible in Section 9.4. It follows that $\Phi_p(x)$ is the minimal polynomial of ζ_p over $\mathbb{Q}$, so that

$$[\mathbb{Q}(\zeta_p) : \mathbb{Q}] = p - 1.$$

We shall see later that in general $[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \varphi(n)$, where $\varphi(n)$ is the Euler phi-function of n (so that $\varphi(p) = p - 1$).

Example: (Splitting Field of $x^p - 2$, p a prime)

Let p be a prime and consider the splitting field of $x^p - 2$. If α is a root of this equation, i.e., $\alpha^p = 2$, then $(\zeta\alpha)^p = 2$ where ζ is any p^{th} root of unity. Hence the solutions of this equation are

$$\zeta \sqrt[p]{2}, \qquad \zeta \text{ a } p^{\text{th}} \text{ root of unity}$$

where as usual the symbol $\sqrt[p]{2}$ denotes the positive real p^{th} root of 2 if we wish to view these elements as complex numbers, and denotes any one solution of $x^p = 2$ if we view these roots abstractly. Since the ratio of the two solutions $\zeta_p \sqrt[p]{2}$ and $\sqrt[p]{2}$ for ζ_p a primitive p^{th} root of unity is just ζ_p, the splitting field of $x^p - 2$ over $\mathbb{Q}$ contains $\mathbb{Q}(\sqrt[p]{2}, \zeta_p)$. On the other hand, all the roots above lie in this field, so that the splitting field is precisely

$$\mathbb{Q}(\sqrt[p]{2}, \zeta_p).$$

This field contains the cyclotomic field of p^{th} roots of unity and is generated over it by $\sqrt[p]{2}$, hence is an extension of degree at most p. It follows that the degree of this extension over $\mathbb{Q}$ is $\leq p(p-1)$. Since both $\mathbb{Q}(\sqrt[p]{2})$ and $\mathbb{Q}(\zeta_p)$ are subfields, the degree of the extension over $\mathbb{Q}$ is divisible by p and by $p - 1$. Since these two numbers are relatively prime it follows that the extension degree is divisible by $p(p-1)$ so that we must have

$$[\mathbb{Q}(\sqrt[p]{2}, \zeta_p) : \mathbb{Q}] = p(p-1)$$

(this is Corollary 22). Note in particular that we have proved $x^p - 2$ remains irreducible over $\mathbb{Q}(\zeta_p)$, which is not at all obvious. We have the following diagram of known subfields:

$$\begin{tikzcd}
& \mathbb{Q}(\sqrt[p]{2}, \zeta_p) \arrow[dl, dash, "p"'] \arrow[dr, dash, "p-1"] & \\
\mathbb{Q}(\zeta_p) \arrow[dr, dash, "p-1"'] & & \mathbb{Q}(\sqrt[p]{2}) \arrow[dl, dash, "p"] \\
& \mathbb{Q} &
\end{tikzcd}$$

The special case $p = 3$ was Example 3 above, where we simply indicated the 3rd roots of unity explicitly.

We now return to the problem of proving it makes no difference how the splitting field of a polynomial $f(x)$ over a field F is constructed. As in Theorem 8 it is convenient to state the result for an arbitrary isomorphism $\varphi : F \xrightarrow{\sim} F'$ between two fields.

Theorem 27. Let $\varphi : F \xrightarrow{\sim} F'$ be an isomorphism of fields. Let $f(x) \in F[x]$ be a polynomial and let $f'(x) \in F'[x]$ be the polynomial obtained by applying φ to the coefficients of $f(x)$. Let E be a splitting field for $f(x)$ over F and let E' be a splitting field for $f'(x)$ over F'. Then the isomorphism φ extends to an isomorphism $\sigma : E \xrightarrow{\sim} E'$, i.e., σ restricted to F is the isomorphism φ:

$$\begin{tikzcd}
\sigma : & E \arrow[r, "\sim"] \arrow[d, dash] & E' \arrow[d, dash] \\
\varphi : & F \arrow[r, "\sim"] & F'
\end{tikzcd}$$

Proof: We shall proceed by induction on the degree n of $f(x)$. As in the discussion before Theorem 8, recall that an isomorphism φ from one field F to another field

F' induces a natural isomorphism between the polynomial rings $F[x]$ and $F'[x]$. In particular, if $f(x)$ and $f'(x)$ correspond to one another under this isomorphism then the irreducible factors of $f(x)$ in $F[x]$ correspond to the irreducible factors of $f'(x)$ in $F'[x]$.

If $f(x)$ has all its roots in F then $f(x)$ splits completely in $F[x]$ and $f'(x)$ splits completely in $F'[x]$ (with its linear factors being the images of the linear factors for $f(x)$). Hence $E = F$ and $E' = F'$, and in this case we may take $\sigma = \varphi$. This shows the result is true for $n = 1$ and in the case where all the irreducible factors of $f(x)$ have degree 1.

Assume now by induction that the theorem has been proved for any field F, isomorphism φ, and polynomial $f(x) \in F[x]$ of degree $< n$. Let $p(x)$ be an irreducible factor of $f(x)$ in $F[x]$ of degree at least 2 and let $p'(x)$ be the corresponding irreducible factor of $f'(x)$ in $F'[x]$. If $\alpha \in E$ is a root of $p(x)$ and $\beta \in E'$ is a root of $p'(x)$, then by Theorem 8 we can extend φ to an isomorphism $\sigma' : F(\alpha) \xrightarrow{\sim} F'(\beta)$:

$$\begin{tikzcd}
\sigma' : & F(\alpha) \arrow[r, "\sim"] \arrow[d, dash] & F'(\beta) \arrow[d, dash] \\
\varphi : & F \arrow[r, "\sim"] & F'.
\end{tikzcd}$$

Let $F_1 = F(\alpha)$, $F_1' = F'(\beta)$, so that we have the isomorphism $\sigma' : F_1 \xrightarrow{\sim} F_1'$. We have $f(x) = (x - \alpha) f_1(x)$ over F_1 where $f_1(x)$ has degree $n - 1$ and $f'(x) = (x - \beta) f_1'(x)$. The field E is a splitting field for $f_1(x)$ over F_1: all the roots of $f_1(x)$ are in E and if they were contained in any smaller extension L containing F_1, then, since F_1 contains α, L would also contain all the roots of $f(x)$, which would contradict the minimality of E as the splitting field of $f(x)$ over F. Similarly E' is a splitting field for $f_1'(x)$ over F_1'. Since the degrees of $f_1(x)$ and $f_1'(x)$ are less than n, by induction there exists a map $\sigma : E \xrightarrow{\sim} E'$ extending the isomorphism $\sigma' : F_1 \xrightarrow{\sim} F_1'$. This gives the extended diagram:

$$\begin{tikzcd}
\sigma : & E \arrow[r, "\sim"] \arrow[d, dash] & E' \arrow[d, dash] \\
\sigma' : & F_1 \arrow[r, "\sim"] \arrow[d, dash] & F_1' \arrow[d, dash] \\
\varphi : & F \arrow[r, "\sim"] & F'.
\end{tikzcd}$$

Then as the diagram indicates, σ restricted to F_1 is the isomorphism σ', so in particular σ restricted to F is σ' restricted to F, which is φ, showing that σ is an extension of φ, completing the proof.

Corollary 28. *(Uniqueness of Splitting Fields)* Any two splitting fields for a polynomial $f(x) \in F[x]$ over a field F are isomorphic.

Proof: Take φ to be the identity mapping from F to itself and E and E' to be two splitting fields for $f(x)(= f'(x))$.

As we mentioned before, this result justifies the terminology of *the* splitting field for $f(x)$ over F, since any two are isomorphic. Splitting fields play a natural role in the study of algebraic elements (if you are adjoining one root of a polynomial, why not adjoin *all* the roots?) and so take a particularly important role in Galois Theory.

We end this section with a discussion of field extensions of F which contain all the roots of *all* polynomials over F.

Definition. The field $\overline{F}$ is called an *algebraic closure* of F if $\overline{F}$ is algebraic over F and if every polynomial $f(x) \in F[x]$ splits completely over $\overline{F}$ (so that $\overline{F}$ can be said to contain all the elements algebraic over F).

Definition. A field K is said to be *algebraically closed* if every polynomial with coefficients in K has a root in K.

It is not obvious that algebraically closed fields exist nor that there exists an algebraic closure of a given field F (we shall prove this shortly).

Note that if K is algebraically closed, then in fact every $f(x) \in K[x]$ has *all* its roots in K, since by definition $f(x)$ has a root $\alpha \in K$, hence has a factor $x - \alpha$ in $K[x]$. The remaining factor of $f(x)$ then is a polynomial in $K[x]$, hence has a root, so has a linear factor etc., so that $f(x)$ must split completely. Hence if K is algebraically closed, then K itself is an algebraic closure of K and the converse is obvious, so that $K = \overline{K}$ if and only if K is algebraically closed.

The next result shows that the process of "taking the algebraic closure" actually stops after one step — taking the algebraic closure of an algebraic closure does not give a larger field: the field is already algebraically closed (notationally: $\overline{\overline{F}} = \overline{F}$).

Proposition 29. Let $\overline{F}$ be an algebraic closure of F. Then $\overline{F}$ is algebraically closed.

Proof: Let $f(x)$ be a polynomial in $\overline{F}[x]$ and let α be a root of $f(x)$. Then α generates an algebraic extension $\overline{F}(\alpha)$ of $\overline{F}$, and $\overline{F}$ is algebraic over F. By Theorem 20, $\overline{F}(\alpha)$ is algebraic over F so in particular its element α is algebraic over F. But then $\alpha \in \overline{F}$, showing $\overline{F}$ is algebraically closed.

Given a field F we have already shown how to construct (finite) extensions of F containing all the roots of any given polynomial $f(x) \in F[x]$. Intuitively, an algebraic closure of F is given by the field "generated" by all of these fields. The difficulty with this is "generated" *where*?, since they are not all subfields of a given field. For a *finite* collection of polynomials $f_1(x), \dots, f_k(x)$, we can identify their splitting fields as subfields of the splitting field of the product polynomial $f_1(x) \cdots f_k(x)$, but the same idea used for an *infinite* number of polynomials requires numerous "bookkeeping" identifications and an application of Zorn's Lemma.

We shall instead construct an algebraic closure of F by first constructing an algebraically closed field containing F. The proof uses a clever idea of Artin which very neatly solves the "bookkeeping" problem of constructing a field containing the appropriate roots of polynomials (which also ultimately relies on Zorn's Lemma) by introducing a separate variable for every polynomial.

Proposition 30. For any field F there exists an algebraically closed field K containing F.

Proof: For every nonconstant monic polynomial $f = f(x)$ with coefficients in F, let x_f denote an indeterminate and consider the polynomial ring $F[\dots, x_f, \dots]$ generated over F by the variables x_f. In this polynomial ring consider the ideal I generated by the polynomials $f(x_f)$. If this ideal is not proper, then 1 is an element of the ideal, hence we have a relation

$$g_1 f_1(x_{f_1}) + g_2 f_2(x_{f_2}) + \dots + g_n f_n(x_{f_n}) = 1$$

where the g_i, $i = 1, 2, \dots, n$, are polynomials in the x_f. For $i = 1, 2, \dots, n$ let $x_{f_i} = x_i$ and let $x_{n+1}, \dots, x_m$ be the remaining variables occurring in the polynomials g_j, $j = 1, 2, \dots, n$. Then the relation above reads

$$g_1(x_1, x_2, \dots, x_m) f_1(x_1) + \dots + g_n(x_1, x_2, \dots, x_m) f_n(x_n) = 1.$$

Let F' be a finite extension of F containing a root α_i of $f_i(x)$ for $i = 1, 2, \dots, n$. Letting $x_i = \alpha_i$, $i = 1, 2, \dots, n$ and setting $x_{n+1} = \dots = x_m = 0$, say, in the polynomial equation above would imply that $0 = 1$ in F', clearly impossible.

Since the ideal I is a proper ideal, it is contained in a maximal ideal $\mathcal{M}$ (this is where Zorn's Lemma is used). Then the quotient

$$K_1 = F[\dots, x_f, \dots]/\mathcal{M}$$

is a field containing (an isomorphic copy of) F. Each of the polynomials f has a root in K_1 by construction, namely the image of x_f, since $f(x_f) \in I \subseteq \mathcal{M}$. We have constructed a field K_1 in which every polynomial with coefficients from F has a root. Performing the same construction with K_1 instead of F gives a field K_2 containing K_1 in which all polynomials with coefficients from K_1 have a root. Continuing in this fashion we obtain a sequence of fields

$$F = K_0 \subseteq K_1 \subseteq K_2 \subseteq \dots \subseteq K_j \subseteq K_{j+1} \subseteq \dots$$

where every polynomial with coefficients in K_j has a root in K_{j+1}, $j = 0, 1, \dots$. Let

$$K = \bigcup_{j \geq 0} K_j$$

be the union of these fields. Then K is clearly a field containing F. Since K is the union of the fields K_j, the coefficients of any polynomial $h(x)$ in $K[x]$ all lie in some field K_N for N sufficiently large. But then $h(x)$ has a root in K_{N+1}, so has a root in K. It follows that K is algebraically closed, completing the proof.

We now use the algebraically closed field containing F to construct an algebraic closure of F:

Proposition 31. Let K be an algebraically closed field and let F be a subfield of K. Then the collection of elements $\overline{F}$ of K that are algebraic over F is an algebraic closure of F. An algebraic closure of F is unique up to isomorphism.

Proof: By definition, $\overline{F}$ is an algebraic extension of F. Every polynomial $f(x) \in F[x]$ splits completely over K into linear factors $x - \alpha$ (the same is true for every polynomial even in $K[x]$). But each α is a root of $f(x)$, so is algebraic over F, hence is an element of $\overline{F}$. It follows that all the linear factors $x - \alpha$ have coefficients in $\overline{F}$, i.e., $f(x)$ splits completely in $\overline{F}[x]$ and $\overline{F}$ is an algebraic closure of F.

The uniqueness (up to isomorphism) of the algebraic closure is natural in light of the uniqueness (up to isomorphism) of splitting fields, and is proved along the same lines together with an application of Zorn's Lemma and will be omitted.

We shall prove later using Galois theory the following result (purely analytic proofs using complex analysis also exist).

Theorem. *(Fundamental Theorem of Algebra)* The field $\mathbb{C}$ is algebraically closed.

By Proposition 31, we immediately obtain:

Corollary 32. The field $\mathbb{C}$ contains an algebraic closure for any of its subfields. In particular, $\overline{\mathbb{Q}}$, the collection of complex numbers algebraic over $\mathbb{Q}$, is an algebraic closure of $\mathbb{Q}$.

The point of these considerations is that all the computations involving elements algebraic over a field F may be viewed as taking place in one (large) field, namely $\overline{F}$. Similarly, we can speak sensibly of the composite of any collection of algebraic extensions by viewing them all as subfields of an algebraic closure. In the case of $\mathbb{Q}$ or finite extensions of $\mathbb{Q}$ we may consider all of our computations as occurring in $\mathbb{C}$.

EXERCISES

1. Determine the splitting field and its degree over $\mathbb{Q}$ for $x^4 - 2$.
2. Determine the splitting field and its degree over $\mathbb{Q}$ for $x^4 + 2$.
3. Determine the splitting field and its degree over $\mathbb{Q}$ for $x^4 + x^2 + 1$.
4. Determine the splitting field and its degree over $\mathbb{Q}$ for $x^6 - 4$.
5. Let K be a finite extension of F. Prove that K is a splitting field over F if and only if every irreducible polynomial in $F[x]$ that has a root in K splits completely in $K[x]$. [Use Theorems 8 and 27.]
6. Let K_1 and K_2 be finite extensions of F contained in the field K, and assume both are splitting fields over F.
   (a) Prove that their composite K_1K_2 is a splitting field over F.
   (b) Prove that $K_1 \cap K_2$ is a splitting field over F. [Use the preceding exercise.]

13.5 SEPARABLE AND INSEPARABLE EXTENSIONS

Let F be a field and let $f(x) \in F[x]$ be a polynomial. Over a splitting field for $f(x)$ we have the factorization

$$f(x) = (x - \alpha_1)^{n_1}(x - \alpha_2)^{n_2} \cdots (x - \alpha_k)^{n_k}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_k$ are distinct elements of the splitting field and $n_i \geq 1$ for all i. Recall that α_i is called a *multiple* root if $n_i > 1$ and is called a *simple* root if $n_i = 1$. The integer n_i is called the *multiplicity* of the root α_i.

Definition. A polynomial over F is called *separable* if it has no multiple roots (i.e., all its roots are distinct). A polynomial which is not separable is called *inseparable*.

Note that if a polynomial $f(x)$ has distinct roots in one splitting field then $f(x)$ has distinct roots in any splitting field (since this is equivalent to $f(x)$ factoring into distinct linear factors, and there is an isomorphism over F between any two splitting fields of $f(x)$ that is bijective on its roots), so that we need not specify the field containing all the roots of $f(x)$.

Examples

(1) The polynomial $x^2 - 2$ is separable over $\mathbb{Q}$ since its two roots $\pm\sqrt{2}$ are distinct. The polynomial $(x^2 - 2)^n$ for any $n \geq 2$ is inseparable since it has the multiple roots $\pm\sqrt{2}$, each with multiplicity n.

(2) The polynomial $x^2 - t$ $(= x^2 + t)$ over the field $F = \mathbb{F}_2(t)$ of rational functions in t with coefficients from $\mathbb{F}_2$ is irreducible as we've seen before, but is not separable. If $\sqrt{t}$ denotes a root in some extension field (note that $\sqrt{t} \notin F$), then

$$(x - \sqrt{t})^2 = x^2 - 2x\sqrt{t} + t = x^2 + t = x^2 - t$$

since F is a field of characteristic 2. Hence this irreducible polynomial has only one root (with multiplicity 2), so is not separable over F.

There is a simple criterion to check whether a polynomial has multiple roots.

Definition. The *derivative* of the polynomial

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in F[x]$$

is defined to be the polynomial

$$D_x f(x) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + 2a_2 x + a_1 \in F[x].$$

This formula is nothing but the usual formula for the derivative of a polynomial familiar from calculus. It is purely algebraic and so can be applied to a polynomial over an arbitrary field F, where the analytic notion of derivative (involving limits — a *continuous* operation) may not exist.

The usual (calculus) formulas for derivatives hold for derivatives in this situation as well, for example the formulas for the derivative of a sum and of a product:

$$D_x(f(x) + g(x)) = D_x f(x) + D_x g(x)$$
$$D_x(f(x)g(x)) = f(x)D_x g(x) + (D_x f(x))g(x).$$

These formulas can be proved directly from the definition for polynomials and do not require any limiting operations and are left as an exercise.

The next proposition shows that the separability of $f(x)$ can be determined by the Euclidean Algorithm in the field where the coefficients of $f(x)$ lie, without passing to a splitting field and factoring $f(x)$.

Proposition 33. A polynomial $f(x)$ has a multiple root α if and only if α is also a root of $D_x f(x)$, i.e., $f(x)$ and $D_x f(x)$ are both divisible by the minimal polynomial for α. In particular, $f(x)$ is separable if and only if it is relatively prime to its derivative: $(f(x), D_x f(x)) = 1$.

Proof: Suppose first that α is a multiple root of $f(x)$. Then over a splitting field,

$$f(x) = (x - \alpha)^n g(x)$$

for some integer $n \geq 2$ and some polynomial $g(x)$. Taking derivatives we obtain

$$D_x f(x) = n(x - \alpha)^{n-1} g(x) + (x - \alpha)^n D_x g(x)$$

which shows ($n \geq 2$) that $D_x f(x)$ has α as a root.

Conversely, suppose that α is a root of both $f(x)$ and $D_x f(x)$. Then write

$$f(x) = (x - \alpha)h(x)$$

for some polynomial $h(x)$ and take the derivative:

$$D_x f(x) = h(x) + (x - \alpha)D_x h(x).$$

Since $D_x f(\alpha) = 0$ by assumption, substituting α into the last equation shows that $h(\alpha) = 0$. Hence $h(x) = (x - \alpha)h_1(x)$ for some polynomial $h_1(x)$, and

$$f(x) = (x - \alpha)^2 h_1(x)$$

showing that α is a multiple root of $f(x)$.

The equivalence with divisibility by the minimal polynomial for α follows from Proposition 9. The last statement is then clear (let α denote any root of a common factor of $f(x)$ and $D_x f(x)$).

Examples

(1) The polynomial $x^{p^n} - x$ over $\mathbb{F}_p$ has derivative $p^n x^{p^n - 1} - 1 = -1$ since the field has characteristic p. Since in this case the derivative has no roots at all, it follows that the polynomial has no multiple roots, hence is separable.

(2) The polynomial $x^n - 1$ has derivative nx^{n-1}. Over any field of characteristic not dividing n (including characteristic 0) this polynomial has only the root 0 (of multiplicity $n - 1$), which is not a root of $x^n - 1$. Hence $x^n - 1$ is separable and there are n distinct n^{th} roots of unity. We saw this directly over $\mathbb{Q}$ by exhibiting n distinct solutions over $\mathbb{C}$.

(3) If F is of characteristic p and p divides n, then there are fewer than n distinct n^{th} roots of unity over F: in this case the derivative is identically 0 since $n = 0$ in F. In fact *every* root of $x^n - 1$ is multiple in this case.

Corollary 34. Every irreducible polynomial over a field of characteristic 0 (for example, $\mathbb{Q}$) is separable. A polynomial over such a field is separable if and only if it is the product of distinct irreducible polynomials.

Proof: Suppose F is a field of characteristic 0 and $p(x) \in F[x]$ is irreducible of degree n. Then the derivative $D_x p(x)$ is a polynomial of degree $n - 1$. Up to constant factors the only factors of $p(x)$ in $F[x]$ are 1 and $p(x)$, so $D_x p(x)$ must be relatively prime to $p(x)$. This shows that any irreducible polynomial over a field of characteristic 0 is separable. The second statement of the corollary is then clear since distinct irreducibles never have zeros in common (by Proposition 9).

The point in the proof of the corollary that can fail in characteristic p is the statement that the derivative $D_x p(x)$ is of degree $n-1$. In characteristic p the derivative of any power x^{pm} of x^p is identically 0:

$$D_x(x^{pm}) = pmx^{pm-1} = 0$$

so it is possible for the degree of the derivative to decrease by more than one. If the derivative $D_x p(x)$ of the *irreducible* polynomial $p(x)$ is nonzero, however, then just as before we conclude that $p(x)$ must be separable.

It is clear from the definition of the derivative that if $p(x)$ is a polynomial whose derivative is 0, then every exponent of x in $p(x)$ must be a multiple of p where p is the characteristic of F:

$$p(x) = a_m x^{mp} + a_{m-1}x^{(m-1)p} + \cdots + a_1 x^p + a_0.$$

Letting

$$p_1(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0$$

we see that $p(x)$ is a polynomial in x^p, namely $p(x) = p_1(x^p)$.

We now prove a simple but important result about raising to the p^{th} power in a field of characteristic p.

Proposition 35. Let F be a field of characteristic p. Then for any $a, b \in F$,

$$(a+b)^p = a^p + b^p, \quad \text{and} \quad (ab)^p = a^p b^p.$$

Put another way, the p^{th}-power map defined by $\varphi(a) = a^p$ is an injective field homomorphism from F to F.

Proof: The Binomial Theorem for expanding $(a+b)^n$ for any positive integer n holds (by the standard induction proof) over any commutative ring:

$$(a+b)^n = a^n + \binom{n}{1}a^{n-1}b + \cdots + \binom{n}{i}a^{n-i}b^i + \cdots + b^n .$$

It should be observed that the binomial coefficients

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}$$

are integers (recall that $m\alpha$ for $m \in \mathbb{Z}$ is defined for α an element of any ring) and here are elements of the prime field.

If p is a prime, then the binomial coefficients $\binom{p}{i}$ for $i = 1, 2, \ldots, p-1$ are all divisible by p since for these values of i the numbers $i!$ and $(p-i)!$ only involve factors smaller than p, hence are relatively prime to p and so cannot cancel the factor of p in the numerator of the expression $\dfrac{p!}{i!(p-i)!}$. It follows that over a field of characteristic p all the intermediate terms in the expansion of $(a+b)^p$ are 0, which gives the first equation of the proposition. The second equation is trivial, as is the fact that φ is injective.

Definition. The map in Proposition 35 is called the *Frobenius endomorphism* of F.

Corollary 36. Suppose that $\mathbb{F}$ is a finite field of characteristic p. Then every element of $\mathbb{F}$ is a p^{th} power in $\mathbb{F}$ (notationally, $\mathbb{F} = \mathbb{F}^p$).

Proof: The injectivity of the Frobenius endomorphism of $\mathbb{F}$ implies that it is also surjective when $\mathbb{F}$ is finite, which is the statement of the corollary.

We now prove the analogue of Corollary 34 for finite fields.

Let $\mathbb{F}$ be a finite field and suppose that $p(x) \in \mathbb{F}[x]$ is an irreducible polynomial with coefficients in $\mathbb{F}$. If $p(x)$ were inseparable then we have seen that $p(x) = q(x^p)$ for some polynomial $q(x) \in \mathbb{F}[x]$. Let

$$q(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0.$$

By Corollary 36, each a_i, $i = 1, 2, \ldots, m$ is a p^{th} power in $\mathbb{F}$, say $a_i = b_i^p$. Then by Proposition 35 we have

$$\begin{aligned} p(x) = q(x^p) &= a_m(x^p)^m + a_{m-1}(x^p)^{m-1} + \cdots + a_1 x^p + a_0 \\ &= b_m^p(x^p)^m + b_{m-1}^p(x^p)^{m-1} + \cdots + b_1^p x^p + b_0^p \\ &= (b_m x^m)^p + (b_{m-1}x^{m-1})^p + \cdots + (b_1 x)^p + (b_0)^p \\ &= (b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0)^p \end{aligned}$$

which shows that $p(x)$ is the p^{th} power of a polynomial in $\mathbb{F}[x]$, a contradiction to the irreducibility of $p(x)$. This proves:

Proposition 37. Every irreducible polynomial over a finite field $\mathbb{F}$ is separable. A polynomial in $\mathbb{F}[x]$ is separable if and only if it is the product of distinct irreducible polynomials in $\mathbb{F}[x]$.

The important part of the proof of this result is the fact that every element in the characteristic p field $\mathbb{F}$ was a p^{th} power in $\mathbb{F}$. This suggests the following definition:

Definition. A field K of characteristic p is called *perfect* if every element of K is a p^{th} power in K, i.e., $K = K^p$. Any field of characteristic 0 is also called perfect.

With this definition, we see that we have proved that every irreducible polynomial over a perfect field is separable. It is not hard to see that if K is not perfect then there are inseparable irreducible polynomials.

Example: (Existence and Uniqueness of Finite Fields)

Let $n > 0$ be any positive integer and consider the splitting field of the polynomial $x^{p^n} - x$ over $\mathbb{F}_p$. We have already seen that this polynomial is separable, hence has precisely p^n roots. Let α and β be any two roots of this polynomial, so that $\alpha^{p^n} = \alpha$ and $\beta^{p^n} = \beta$. Then $(\alpha\beta)^{p^n} = \alpha\beta$, $(\alpha^{-1})^{p^n} = \alpha^{-1}$ and by Proposition 35 also

$$(\alpha + \beta)^{p^n} = \alpha^{p^n} + \beta^{p^n} = \alpha + \beta.$$

Hence the set $\mathbb{F}$ consisting of the p^n distinct roots of $x^{p^n} - x$ over $\mathbb{F}_p$ is *closed* under addition, multiplication and inverses in its splitting field. It follows that $\mathbb{F}$ is a subfield, hence in fact must be the splitting field. Since the number of elements is p^n, we have $[\mathbb{F} : \mathbb{F}_p] = n$, which shows that there exist finite fields of degree n over $\mathbb{F}_p$ for any $n > 0$.

Let now $\mathbb{F}$ be any finite field of characteristic p. If $\mathbb{F}$ is of dimension n over its prime subfield $\mathbb{F}_p$, then $\mathbb{F}$ has precisely p^n elements. Since the multiplicative group $\mathbb{F}^\times$ is (in fact cyclic) of order $p^n - 1$, we have $\alpha^{p^n-1} = 1$ for every $\alpha \neq 0$ in $\mathbb{F}$, so that $\alpha^{p^n} = \alpha$ for every $\alpha \in \mathbb{F}$. But this means α is a root of $x^{p^n} - x$, hence $\mathbb{F}$ is contained in a splitting field for this polynomial. Since we have seen that the splitting field has order p^n this shows that $\mathbb{F}$ is a splitting field for $x^{p^n} - x$. Since splitting fields are unique up to isomorphism, this proves that *finite fields of any order p^n exist and are unique up to isomorphism.* We shall denote the finite field of order p^n by $\mathbb{F}_{p^n}$.

We shall consider finite fields more later.

We now investigate further the structure of inseparable irreducible polynomials over fields of characteristic p. We have seen above that if $p(x)$ is an irreducible polynomial which is not separable, then its derivative $D_x p(x)$ is identically 0, so that $p(x) = p_1(x^p)$ for some polynomial $p_1(x)$. The polynomial $p_1(x)$ may or may not itself be separable. If not, then it too is a polynomial in x^p, $p_1(x) = p_2(x^p)$, so that $p(x)$ is a polynomial in x^{p^2}: $p(x) = p_2(x^{p^2})$. Continuing in this fashion we see that there is a uniquely defined power p^k of p such that $p(x) = p_k(x^{p^k})$ where $p_k(x)$ has nonzero derivative. It is clear that $p_k(x)$ is irreducible since any factorization of $p_k(x)$ would, after replacing x by x^{p^k}, immediately imply a factorization of the irreducible $p(x)$. It follows that $p_k(x)$ is separable. We summarize this as:

Proposition 38. Let $p(x)$ be an irreducible polynomial over a field F of characteristic p. Then there is a unique integer $k \geq 0$ and a unique irreducible separable polynomial $p_{sep}(x) \in F[x]$ such that

$$p(x) = p_{sep}(x^{p^k}).$$

Definition. Let $p(x)$ be an irreducible polynomial over a field of characteristic p. The degree of $p_{sep}(x)$ in the last proposition is called the *separable degree* of $p(x)$, denoted $\deg_s p(x)$. The integer p^k in the proposition is called the *inseparable degree* of $p(x)$, denoted $\deg_i p(x)$.

From the definitions and the proposition we see that $p(x)$ is separable if and only if its inseparability degree is 1 if and only if its degree is equal to its separable degree. Also, computing degrees in the relation $p(x) = p_{sep}(x^{p^k})$ we see that

$$\deg p(x) = \deg_s p(x) \deg_i p(x).$$

Examples

(1) The polynomial $p(x) = x^2 - t$ over $F = \mathbb{F}_2(t)$ considered above has derivative 0, hence is not separable (as we determined earlier). Here $p_{sep}(x) = x - t$ with inseparability degree 2.

(2) The polynomial $p(x) = x^{2^m} - t$ over $F = \mathbb{F}_2(t)$ is irreducible with the same separable polynomial part, but with inseparability degree 2^m.

(3) The polynomial $(x^{p^2} - t)(x^p - t)$ over $F = \mathbb{F}_p(t)$ has (two) inseparable irreducible factors so is inseparable. This polynomial cannot be written in the form $f_{sep}(x^{p^k})$ where $f_{sep}(x)$ is separable, which is the reason we restricted to *irreducible* polynomials above. This example also shows that there is no analogous factorization to define the separable and inseparable degrees of a general polynomial.

The notion of separability carries over to the fields generated by the roots of these polynomials.

Definition. The field K is said to be *separable* (or *separably algebraic*) over F if every element of K is the root of a separable polynomial over F (equivalently, the minimal polynomial over F of every element of K is separable). A field which is not separable is *inseparable.*

We have seen that the issue of separability is straightforward for finite extensions of perfect fields since for these fields the minimal polynomial of an algebraic element is irreducible hence separable.

Corollary 39. Every finite extension of a perfect field is separable. In particular, every finite extension of either $\mathbb{Q}$ or a finite field is separable.

We shall consider separable and inseparable extensions more after developing some Galois Theory, in particular defining the separable and inseparable *degree* of the extension K/F.

EXERCISES

1. Prove that the derivative D_x of a polynomial satisfies $D_x(f(x) + g(x)) = D_x(f(x)) + D_x(g(x))$ and $D_x(f(x)g(x)) = D_x(f(x))g(x) + D_x(g(x))f(x)$ for any two polynomials $f(x)$ and $g(x)$.
2. Find all irreducible polynomials of degrees 1, 2 and 4 over $\mathbb{F}_2$ and prove that their product is $x^{16} - x$.
3. Prove that d divides n if and only if $x^d - 1$ divides $x^n - 1$. [Note that if $n = qd + r$ then $x^n - 1 = (x^{qd+r} - x^r) + (x^r - 1)$.]
4. Let $a > 1$ be an integer. Prove for any positive integers n, d that d divides n if and only if $a^d - 1$ divides $a^n - 1$ (cf. the previous exercise). Conclude in particular that $\mathbb{F}_{p^d} \subseteq \mathbb{F}_{p^n}$ if and only if d divides n.
5. For any prime p and any nonzero $a \in \mathbb{F}_p$ prove that $x^p - x + a$ is irreducible and separable over $\mathbb{F}_p$. [For the irreducibility: One approach — prove first that if α is a root then $\alpha + 1$ is also a root. Another approach — suppose it's reducible and compute derivatives.]
6. Prove that $x^{p^n - 1} - 1 = \prod_{\alpha \in \mathbb{F}_{p^n}^\times}(x - \alpha)$. Conclude that $\prod_{\alpha \in \mathbb{F}_{p^n}^\times} \alpha = (-1)^{p^n}$ so the product of the nonzero elements of a finite field is $+1$ if $p = 2$ and -1 if p is odd. For p odd and $n = 1$ derive *Wilson's Theorem*: $(p-1)! \equiv -1 \pmod{p}$.

7. Suppose K is a field of characteristic p which is not a perfect field: $K \neq K^p$. Prove there exist irreducible inseparable polynomials over K. Conclude that there exist inseparable finite extensions of K.

8. Prove that $f(x)^p = f(x^p)$ for any polynomial $f(x) \in \mathbb{F}_p[x]$.

9. Show that the binomial coefficient $\binom{pn}{pi}$ is the coefficient of x^{pi} in the expansion of $(1+x)^{pn}$. Working over $\mathbb{F}_p$ show that this is the coefficient of $(x^p)^i$ in $(1 + x^p)^n$ and hence prove that $\binom{pn}{pi} \equiv \binom{n}{i} \pmod{p}$.

10. Let $f(x_1, x_2, \dots, x_n) \in \mathbb{Z}[x_1, x_2, \dots, x_n]$ be a polynomial in the variables $x_1, x_2, \dots, x_n$ with integer coefficients. For any prime p prove that the polynomial
$$f(x_1, x_2, \dots, x_n)^p - f(x_1^p, x_2^p, \dots, x_n^p) \in \mathbb{Z}[x_1, x_2, \dots, x_n]$$
has all its coefficients divisible by p.

11. Suppose $K[x]$ is a polynomial ring over the field K and F is a subfield of K. If F is a perfect field and $f(x) \in F[x]$ has no repeated irreducible factors in $F[x]$, prove that $f(x)$ has no repeated irreducible factors in $K[x]$.

13.6 CYCLOTOMIC POLYNOMIALS AND EXTENSIONS

The purpose of this section is to prove that the cyclotomic extension
$$\mathbb{Q}(\zeta_n)/\mathbb{Q}$$
generated by the n^{th} roots of unity over $\mathbb{Q}$ introduced in Section 4 is of degree $\varphi(n)$ where φ denotes Euler's phi-function (= the number of integers a, $1 \leq a < n$ relatively prime to n = the order of the group $(\mathbb{Z}/n\mathbb{Z})^\times$).

Definition. Let μ_n denote the *group of n^{th} roots of unity over* $\mathbb{Q}$.

Then as we have already observed, $\mathbb{Z}/n\mathbb{Z} \cong \mu_n$ as groups (under multiplication on the right, addition on the left), given explicitly by the map $a \mapsto (\zeta_n)^a$ for a fixed primitive n^{th} root of unity. The primitive n^{th} roots of unity are given by the residue classes prime to n so there are precisely $\varphi(n)$ primitive n^{th} roots of unity.

If d is a divisor of n and ζ is a d^{th} root of unity, then ζ is also an n^{th} root of unity since $\zeta^n = (\zeta^d)^{n/d} = 1$. Hence
$$\mu_d \subseteq \mu_n \qquad \text{for all } d \mid n.$$
Conversely, the order of any element of the group μ_n is a divisor of n so that if ζ is an n^{th} root of unity which is also a d^{th} root of unity for some smaller d then $d \mid n$.

Definition. Define the n^{th} *cyclotomic polynomial* $\Phi_n(x)$ to be the polynomial whose roots are the primitive n^{th} roots of unity:
$$\Phi_n(x) = \prod_{\zeta \text{ primitive } \in \mu_n} (x - \zeta) = \prod_{\substack{1 \leq a < n \\ (a,n)=1}} (x - {\zeta_n}^a)$$
(which is of degree $\varphi(n)$).

The roots of the polynomial $x^n - 1$ are precisely the n^{th} roots of unity so we have the factorization

$$x^n - 1 = \prod_{\substack{\zeta^n=1 \\ \text{i.e. } \zeta\in\mu_n}} (x - \zeta).$$

If we group together the factors $(x - \zeta)$ where ζ is an element of order d in μ_n (i.e., ζ is a primitive d^{th} root of unity) we obtain

$$x^n - 1 = \prod_{d|n} \prod_{\substack{\zeta\in\mu_d \\ \zeta \text{ primitive}}} (x - \zeta).$$

The inner product is $\Phi_d(x)$ by definition so we have the factorization

$$x^n - 1 = \prod_{d|n} \Phi_d(x). \tag{13.4}$$

Note incidentally that comparing degrees gives the identity

$$n = \sum_{d|n} \varphi(d).$$

This factorization allows us to compute $\Phi_n(x)$ for any n recursively: clearly $\Phi_1(x) = x - 1$ and $\Phi_2(x) = x + 1$. Then

$$x^3 - 1 = \Phi_1(x)\Phi_3(x) = (x - 1)\Phi_3(x)$$

which gives

$$\Phi_3(x) = x^2 + x + 1.$$

Similarly

$$x^4 - 1 = \Phi_1(x)\Phi_2(x)\Phi_4(x) = (x - 1)(x + 1)\Phi_4(x)$$

gives

$$\Phi_4(x) = x^2 + 1$$

(in these cases these could also be obtained directly from the explicit roots of unity). Continuing in this fashion we can compute $\Phi_n(x)$ for any n. Note also that for p a prime we recover our polynomial

$$\Phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1.$$

For some small values of n the polynomials are

$$\begin{aligned}
\Phi_5(x) &= x^4 + x^3 + x^2 + x + 1 \\
\Phi_6(x) &= x^2 - x + 1 \\
\Phi_7(x) &= x^6 + x^5 + x^4 + x^3 + x^2 + x + 1 \\
\Phi_8(x) &= x^4 + 1 \\
\Phi_9(x) &= x^6 + x^3 + 1 \\
\Phi_{10}(x) &= x^4 - x^3 + x^2 - x + 1 \\
\Phi_{11}(x) &= x^{10} + x^9 + \cdots + x + 1 \\
\Phi_{12}(x) &= x^4 - x^2 + 1.
\end{aligned}$$

For all the values computed above, $\Phi_n(x)$ was a (monic) polynomial with integer coefficients. This is always the case:

Lemma 40. The cyclotomic polynomial $\Phi_n(x)$ is a monic polynomial in $\mathbb{Z}[x]$ of degree $\varphi(n)$.

Proof: It is clear that $\Phi_n(x)$ is monic and has degree $\varphi(n)$. We must show the coefficients lie in $\mathbb{Z}$. We use induction on n. The result is true for $n = 1$ (and $n \leq 12$). Assume by induction that $\Phi_d(x) \in \mathbb{Z}[x]$ for all $1 \leq d < n$. Then $x^n - 1 = f(x)\Phi_n(x)$ where $f(x) = \prod_{\substack{d|n \\ d<n}} \Phi_d(x)$ is monic and has coefficients in $\mathbb{Z}$. Since $f(x)$ clearly divides $x^n - 1$ in $F[x]$ where $F = \mathbb{Q}(\zeta_n)$ is the field of n^{th} roots of unity and both $f(x)$ and $x^n - 1$ have coefficients in $\mathbb{Q}$, $f(x)$ divides $x^n - 1$ in $\mathbb{Q}[x]$ by the Division Algorithm (cf. the remark at the end of Section 9.2). By Gauss' Lemma, $f(x)$ divides $x^n - 1$ in $\mathbb{Z}[x]$, hence $\Phi_n(x) \in \mathbb{Z}[x]$.

We remark in passing that while all the coefficients of $\Phi_n(x)$ in the examples computed above were 0, ±1, it is known that there are cyclotomic polynomials with arbitrarily large coefficients.

Theorem 41. The cyclotomic polynomial $\Phi_n(x)$ is an irreducible monic polynomial in $\mathbb{Z}[x]$ of degree $\varphi(n)$.

Proof: We must show that $\Phi_n(x)$ is irreducible. If not then we have a factorization

$$\Phi_n(x) = f(x)g(x) \qquad \text{with } f(x), g(x) \text{ monic in } \mathbb{Z}[x]$$

where we take $f(x)$ to be an *irreducible* factor of $\Phi_n(x)$. Let ζ be a primitive n^{th} root of 1 which is a root of $f(x)$ (so then $f(x)$ is the minimal polynomial for ζ over $\mathbb{Q}$) and let p denote *any* prime not dividing n. Then ζ^p is again a primitive n^{th} root of 1, hence is a root of either $f(x)$ or $g(x)$.

Suppose $g(\zeta^p) = 0$. Then ζ is a root of $g(x^p)$ and since $f(x)$ is the minimal polynomial for ζ, $f(x)$ must divide $g(x^p)$ in $\mathbb{Z}[x]$, say

$$g(x^p) = f(x)h(x)\,, \qquad h(x) \in \mathbb{Z}[x].$$

If we reduce this equation mod p, we obtain

$$\bar{g}(x^p) = \bar{f}(x)\bar{h}(x) \qquad \text{in } \mathbb{F}_p[x].$$

By the remarks of the last section,

$$\bar{g}(x^p) = (\bar{g}(x))^p$$

so we have the equation

$$(\bar{g}(x))^p = \bar{f}(x)\bar{h}(x)$$

in the U.F.D. $\mathbb{F}_p[x]$. It follows that $\bar{f}(x)$ and $\bar{g}(x)$ have a factor in common in $\mathbb{F}_p[x]$.

Now, from $\Phi_n(x) = f(x)g(x)$ we see by reducing mod p that $\overline{\Phi}_n(x) = \bar{f}(x)\bar{g}(x)$, and so by the above it follows that $\overline{\Phi}_n(x) \in \mathbb{F}_p[x]$ has a multiple root. But then also $x^n - 1$ would have a multiple root over $\mathbb{F}_p$ since it has $\overline{\Phi}_n(x)$ as a factor. This is a contradiction since we have seen in the last section that there are n distinct roots of $x^n - 1$ over any field of characteristic not dividing n.

Hence ζ^p must be a root of $f(x)$. Since this applies to every root ζ of $f(x)$, it follows that ζ^a is a root of $f(x)$ for every integer a relatively prime to n: write $a = p_1 p_2 \cdots p_k$ as a product of (not necessarily distinct) primes not dividing n so that ζ^{p_1} is a root of $f(x)$, so also $(\zeta^{p_1})^{p_2}$ is a root of $f(x)$, etc. But this means that *every* primitive n^{th} root of unity is a root of $f(x)$, i.e., $f(x) = \Phi_n(x)$, showing $\Phi_n(x)$ is irreducible.

Corollary 42. The degree over $\mathbb{Q}$ of the cyclotomic field of n^{th} roots of unity is $\varphi(n)$:

$$[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \varphi(n).$$

Proof: By the theorem, $\Phi_n(x)$ is the minimal polynomial for any primitive n^{th} root of unity ζ_n.

Example

The cyclotomic field $\mathbb{Q}(\zeta_8)$ of the 8^{th} roots of unity is of degree $\varphi(8) = 4$ over $\mathbb{Q}$. This field contains the 4^{th} roots of unity, i.e., $\mathbb{Q}(i) \subset \mathbb{Q}(\zeta_8)$ as well as the element $\zeta_8 + \zeta_8{}^7 = \sqrt{2}$ (recall the explicit roots of unity in Section 4). It follows that

$$\mathbb{Q}(\zeta_8) = \mathbb{Q}(i, \sqrt{2}).$$

One interesting number-theoretic application of the cyclotomic polynomials outlined in the exercises is the proof that for any n there are infinitely many primes which are congruent to 1 modulo n. The complete factorization in $\mathbb{F}_p[x]$ of $\Phi_\ell(x)$ for a prime ℓ (which is irreducible in $\mathbb{Z}[x]$) is described in Exercise 8 below.

We shall return to the example of cyclotomic fields after we have developed some Galois Theory.

EXERCISES

1. Suppose m and n are relatively prime positive integers. Let ζ_m be a primitive m^{th} root of unity and let ζ_n be a primitive n^{th} root of unity. Prove that $\zeta_m \zeta_n$ is a primitive mn^{th} root of unity.
2. Let ζ_n be a primitive n^{th} root of unity and let d be a divisor of n. Prove that ζ_n^d is a primitive $(n/d)^{\text{th}}$ root of unity.
3. Prove that if a field contains the n^{th} roots of unity for n odd then it also contains the $2n^{\text{th}}$ roots of unity.
4. Prove that if $n = p^k m$ where p is a prime and m is relatively prime to p then there are precisely m distinct n^{th} roots of unity over a field of characteristic p.
5. Prove there are only a finite number of roots of unity in any finite extension K of $\mathbb{Q}$.
6. Prove that for n odd, $n > 1$, $\Phi_{2n}(x) = \Phi_n(-x)$.
7. Use the Möbius Inversion formula indicated in Section 14.3 to prove

$$\Phi_m(x) = \prod_{d|n} (x^d - 1)^{\mu(m/d)}.$$

8. Let ℓ be a prime and let $\Phi_\ell(x) = \frac{x^\ell - 1}{x-1} = x^{\ell-1} + x^{\ell-2} + \ldots + x + 1 \in \mathbb{Z}[x]$ be the ℓ^{th} cyclotomic polynomial, which is irreducible over $\mathbb{Z}$ by Theorem 41. This exercise determines the factorization of $\Phi_\ell(x)$ modulo p for any prime p. Let ζ denote any fixed primitive ℓ^{th} root of unity.

(a) Show that if $p = l$ then $\Phi_\ell(x) = (x-1)^{\ell-1} \in \mathbb{F}_\ell[x]$.

(b) Suppose $p \neq \ell$ and let f denote the order of p mod ℓ, i.e., f is the smallest power of p with $p^f \equiv 1 \bmod \ell$. Use the fact that $\mathbb{F}_{p^n}^\times$ is a cyclic group to show that $n = f$ is the smallest power p^n of p with $\zeta \in \mathbb{F}_{p^n}$. Conclude that the minimal polynomial of ζ over $\mathbb{F}_p$ has degree f.

(c) Show that $\mathbb{F}_p(\zeta) = \mathbb{F}_p(\zeta^a)$ for any integer a not divisible by ℓ. [One inclusion is obvious. For the other, note that $\zeta = (\zeta^a)^b$ where b is the multiplicative inverse of a mod ℓ.] Conclude using (b) that, in $\mathbb{F}_p[x]$, $\Phi_\ell(x)$ is the product of $\frac{\ell-1}{f}$ distinct irreducible polynomials of degree f.

(d) In particular, prove that, viewed in $\mathbb{F}_p[x]$, $\Phi_7(x) = x^6 + x^5 + \ldots + x + 1$ is $(x-1)^6$ for $p = 7$, a product of distinct linear factors for $p \equiv 1 \bmod 7$, a product of 3 irreducible quadratics for $p \equiv 6 \bmod 7$, a product of 2 irreducible cubics for $p \equiv 2, 4 \bmod 7$, and is irreducible for $p \equiv 3, 5 \bmod 7$.

9. Suppose A is an $n \times n$ matrix over $\mathbb{C}$ for which $A^k = I$ for some integer $k \geq 1$. Show that A can be diagonalized. Show that the matrix $A = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}$ where α is an element of a field of characteristic p satisfies $A^p = I$ and cannot be diagonalized if $\alpha \neq 0$.

10. Let φ denote the Frobenius map $x \mapsto x^p$ on the finite field $\mathbb{F}_{p^n}$. Prove that φ gives an isomorphism of $\mathbb{F}_{p^n}$ to itself (such an isomorphism is called an *automorphism*). Prove that φ^n is the identity map and that no lower power of φ is the identity.

11. Let φ denote the Frobenius map $x \mapsto x^p$ on the finite field $\mathbb{F}_{p^n}$ as in the previous exercise. Determine the rational canonical form over $\mathbb{F}_p$ for φ considered as an $\mathbb{F}_p$-linear transformation of the n-dimensional $\mathbb{F}_p$-vector space $\mathbb{F}_{p^n}$.

12. Let φ denote the Frobenius map $x \mapsto x^p$ on the finite field $\mathbb{F}_{p^n}$ as in the previous exercise. Determine the Jordan canonical form (over a field containing all the eigenvalues) for φ considered as an $\mathbb{F}_p$-linear transformation of the n-dimensional $\mathbb{F}_p$-vector space $\mathbb{F}_{p^n}$.

13. (*Wedderburn's Theorem on Finite Division Rings*) This exercise outlines a proof (following Witt) of Wedderburn's Theorem that a finite division ring D is a field (i.e., is commutative).

(a) Let Z denote the center of D (i.e., the elements of D which commute with every element of D). Prove that Z is a field containing $\mathbb{F}_p$ for some prime p. If $Z = \mathbb{F}_q$ prove that D has order q^n for some integer n [D is a vector space over Z].

(b) The nonzero elements $D^\times$ of D form a multiplicative group. For any $x \in D^\times$ show that the elements of D which commute with x form a division ring which contains Z. Show that this division ring is of order q^m for some integer m and that $m < n$ if x is not an element of Z.

(c) Show that the class equation (Theorem 4.7) for the group $D^\times$ is

$$q^n - 1 = (q-1) + \sum_{i=1}^{r} \frac{q^n - 1}{|C_{D^\times}(x_i)|}$$

where $x_1, x_2, \ldots, x_r$ are representatives of the distinct conjugacy classes in $D^\times$ not contained in the center of $D^\times$. Conclude from (b) that for each i, $|C_{D^\times}(x_i)| = q^{m_i} - 1$ for some $m_i < n$.

(d) Prove that since $\dfrac{q^n - 1}{q^{m_i} - 1}$ is an integer (namely, the index $|D^\times : C_{D^\times}(x_i)|$) then m_i divides n (cf. Exercise 4 of Section 5). Conclude that $\Phi_n(x)$ divides $(x^n - 1)/(x^{m_i} - 1)$ and hence that the integer $\Phi_n(q)$ divides $(q^n - 1)/(q^{m_i} - 1)$ for $i = 1, 2, \dots, r$.

(e) Prove that (c) and (d) imply that $\Phi_n(q) = \prod_{\zeta \text{ primitive}}(q - \zeta)$ divides $q - 1$. Prove that $|q - \zeta| > q - 1$ (complex absolute value) for any root of unity $\zeta \neq 1$ [note that 1 is the closest point on the unit circle in $\mathbb{C}$ to the point q on the real line]. Conclude that $n = 1$, i.e., that $D = Z$ is a field.

The following exercises provide a proof that for any positive integer m there are infinitely many primes p with $p \equiv 1 \pmod{m}$. This is a special case of *Dirichlet's Theorem on Primes in Arithmetic Progressions* which states more generally that there are infinitely many primes p with $p \equiv a \pmod{m}$ for any a relatively prime to m.

14. Given any monic polynomial $P(x) \in \mathbb{Z}[x]$ of degree at least one show that there are infinitely many distinct prime divisors of the integers

$$P(1), P(2), P(3), \dots, P(n), \dots.$$

[Suppose $p_1, p_2, \dots, p_k$ are the only primes dividing the values $P(n)$, $n = 1, 2, \dots$. Let N be an integer with $P(N) = a \neq 0$. Show that $Q(x) = a^{-1}P(N + a\, p_1 p_2 \dots p_k\, x)$ is an element of $\mathbb{Z}[x]$ and that $Q(n) \equiv 1 \pmod{p_1 p_2 \dots p_k}$ for $n = 1, 2, \dots$. Conclude that there is some integer M such that $Q(M)$ has a prime factor different from $p_1, p_2, \dots, p_k$ and hence that $P(N + ap_1 p_2 \cdots p_k M)$ has a prime factor different from $p_1, p_2, \dots, p_k$.]

15. Let p be an odd prime not dividing m and let $\Phi_m(x)$ be the m^{th} cyclotomic polynomial. Suppose $a \in \mathbb{Z}$ satisfies $\Phi_m(a) \equiv 0 \pmod{p}$. Prove that a is relatively prime to p and that the order of a in $(\mathbb{Z}/p\mathbb{Z})^\times$ is precisely m. [Since

$$x^m - 1 = \prod_{d \mid m} \Phi_d(x) = \Phi_m(x) \prod_{\substack{d \mid m \\ d < m}} \Phi_d(x)$$

we see first that $a^m - 1 \equiv 0 \pmod{p}$ i.e., $a^m \equiv 1 \pmod{p}$. If the order of a mod p were less than m, then $a^d \equiv 1 \pmod{p}$ for some d dividing m, so then $\Phi_d(a) \equiv 0 \pmod{p}$ for some $d < m$. But then $x^m - 1$ would have a as a multiple root mod p, a contradiction.]

16. Let $a \in \mathbb{Z}$. Show that if p is an odd prime dividing $\Phi_m(a)$ then either p divides m or $p \equiv 1 \pmod{m}$.

17. Prove there are infinitely many primes p with $p \equiv 1 \pmod{m}$.

CHAPTER 14

Galois Theory

14.1 BASIC DEFINITIONS

In the previous chapter we proved the existence of a finite extension of a field F which contains all the roots of a given polynomial $f(x)$ whose coefficients are in F. The main idea of Galois Theory (named for Évariste Galois, 1811–1832) is to consider the relation of the group of permutations of the roots of $f(x)$ to the algebraic structure of its splitting field. The connection is given by the Fundamental Theorem of the next section. It can be viewed as another (extremely elegant) application of the important idea in mathematics that one (in our case algebraic) object *acting* on another provides structural information about both.

In this section we introduce the terminology and basic properties of the objects of interest. Let K be a field.

Definition.

(1) An isomorphism σ of K with itself is called an *automorphism* of K. The collection of automorphisms of K is denoted $\text{Aut}(K)$. If $\alpha \in K$ we shall write $\sigma\alpha$ for $\sigma(\alpha)$.

(2) An automorphism $\sigma \in \text{Aut}(K)$ is said to *fix* an element $\alpha \in K$ if $\sigma\alpha = \alpha$. If F is a subset of K (for example, a subfield), then an automorphism σ is said to *fix* F if it fixes all the elements of F, i.e., $\sigma a = a$ for all $a \in F$.

Note that any field has at least one automorphism, the identity map, denoted by 1 and sometimes called the *trivial* automorphism.

The prime field of K is generated by $1 \in K$ and since any automorphism σ takes 1 to 1 (and 0 to 0), i.e., $\sigma(1) = 1$, it follows that $\sigma a = a$ for all a in the prime field. Hence any automorphism of a field K fixes its prime subfield. In particular we see that $\mathbb{Q}$ and $\mathbb{F}_p$ have only the trivial automorphism: $\text{Aut}(\mathbb{Q}) = \{1\}$ and $\text{Aut}(\mathbb{F}_p) = \{1\}$.

Definition. Let K/F be an extension of fields. Let $\text{Aut}(K/F)$ be the collection of automorphisms of K which fix F.

Note that if F is the prime subfield of K then $\text{Aut}(K) = \text{Aut}(K/F)$ since every automorphism of K automatically fixes F.

If σ and τ are automorphisms of K then the composite $\sigma\tau$ (and also the composite $\tau\sigma$, which may not be the same) is defined and is again an automorphism of K.

Proposition 1. $\text{Aut}(K)$ is a group under composition and $\text{Aut}(K/F)$ is a subgroup.

Proof: It is clear that $\text{Aut}(K)$ is a group. If σ and τ are automorphisms of K which fix F then also $\sigma\tau$ and σ^{-1} are the identity on F, which shows that $\text{Aut}(K/F)$ is a subgroup.

The following proposition is extremely useful for determining the automorphisms of algebraic extensions.

Proposition 2. Let K/F be a field extension and let $\alpha \in K$ be algebraic over F. Then for any $\sigma \in \text{Aut}(K/F)$, $\sigma\alpha$ is a root of the minimal polynomial for α over F i.e., $\text{Aut}(K/F)$ permutes the roots of irreducible polynomials. Equivalently, any polynomial with coefficients in F having α as a root also has $\sigma\alpha$ as a root.

Proof: Suppose α satisfies the equation

$$\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_1\alpha + a_0 = 0$$

where $a_0, a_1, \ldots, a_{n-1}$ are elements of F. Applying the automorphism σ we obtain (using the fact that σ is an additive homomorphism)

$$\sigma(\alpha^n) + \sigma(a_{n-1}\alpha^{n-1}) + \cdots + \sigma(a_1\alpha) + \sigma(a_0) = \sigma(0) = 0.$$

Using the fact that σ is also a multiplicative homomorphism this becomes

$$(\sigma(\alpha))^n + \sigma(a_{n-1})(\sigma(\alpha))^{n-1} + \cdots + \sigma(a_1)(\sigma(\alpha)) + \sigma(a_0) = 0.$$

By assumption, σ fixes all the elements of F, so $\sigma(a_i) = a_i, i = 0, 1, \ldots, n-1$. Hence

$$(\sigma\alpha)^n + a_{n-1}(\sigma\alpha)^{n-1} + \cdots + a_1(\sigma\alpha) + a_0 = 0.$$

But this says precisely that $\sigma\alpha$ is a root of the same polynomial over F as α. This proves the proposition.

Examples

(1) Let $K = \mathbb{Q}(\sqrt{2})$. If $\tau \in \text{Aut}(\mathbb{Q}(\sqrt{2})) = \text{Aut}(\mathbb{Q}(\sqrt{2})/\mathbb{Q})$, then $\tau(\sqrt{2}) = \pm\sqrt{2}$ since these are the two roots of the minimal polynomial for $\sqrt{2}$. Since τ fixes $\mathbb{Q}$, this determines τ completely:

$$\tau(a + b\sqrt{2}) = a \pm b\sqrt{2}.$$

The map $\sqrt{2} \mapsto \sqrt{2}$ is just the identity automorphism 1 of $\mathbb{Q}(\sqrt{2})$. The map $\sigma : \sqrt{2} \mapsto -\sqrt{2}$ is the isomorphism considered in Example 2 following Corollary 13.7. Hence $\text{Aut}(\mathbb{Q}(\sqrt{2})) = \text{Aut}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}) = \{1, \sigma\}$ is a cyclic group of order 2 generated by σ.

(2) Let $K = \mathbb{Q}(\sqrt[3]{2})$. As before, if $\tau \in \text{Aut}(K/\mathbb{Q})$, then τ is completely determined by its action on $\sqrt[3]{2}$ since

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

Since $\tau\sqrt[3]{2}$ must be a root of $x^3 - 2$ and the other two roots of this equation are not elements of K (recall the splitting field of this polynomial is degree 6 over $\mathbb{Q}$), the only possibility is $\tau\sqrt[3]{2} = \sqrt[3]{2}$ i.e., $\tau = 1$. Hence $\text{Aut}(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}) = 1$ is the trivial group.

In general, if K is generated over F by some collection of elements, then any automorphism $\sigma \in \text{Aut}(K/F)$ is completely determined by what it does to the generators. If K/F is finite then K is finitely generated over F by algebraic elements so by the proposition the number of automorphisms of K fixing F is finite, i.e., $\text{Aut}(K/F)$ is a finite group. In particular, the automorphisms of a finite extension can be considered as permutations of the roots of a finite number of equations (not every permutation gives rise to an automorphism, however, as Example 2 above illustrates). It was the investigation of permutations of the roots of equations that led Galois to the theory we are describing.

We have associated to each field extension K/F (equivalently, with a subfield F of K) a *group*, $\text{Aut}(K/F)$, the group of automorphisms of K which fix F. One can also reverse this process and associate to each group of automorphisms a field extension.

Proposition 3. Let $H \leq \text{Aut}(K)$ be a subgroup of the group of automorphisms of K. Then the collection F of elements of K fixed by all the elements of H is a subfield of K.

Proof: Let $h \in H$ and let $a, b \in F$. Then by definition $h(a) = a$, $h(b) = b$ so that $h(a \pm b) = h(a) \pm h(b) = a \pm b$, $h(ab) = h(a)h(b) = ab$ and $h(a^{-1}) = h(a)^{-1} = a^{-1}$, so that F is closed, hence a subfield of K.

Note that it is not important in this proposition that H actually be a *subgroup* of $\text{Aut}(K)$ — the collection of elements of K fixed by all the elements of a *subset* of $\text{Aut}(K)$ is also a subfield of K.

Definition. If H is a subgroup of the group of automorphisms of K, the subfield of K fixed by all the elements of H is called the *fixed field* of H.

Proposition 4. The association of groups to fields and fields to groups defined above is inclusion reversing, namely

(1) if $F_1 \subseteq F_2 \subseteq K$ are two subfields of K then $\text{Aut}(K/F_2) \leq \text{Aut}(K/F_1)$, and

(2) if $H_1 \leq H_2 \leq \text{Aut}(K)$ are two subgroups of automorphisms with associated fixed fields F_1 and F_2, respectively, then $F_2 \subseteq F_1$.

Proof: Any automorphism of K that fixes F_2 also fixes its subfield F_1, which gives (1). The second assertion is proved similarly.

Examples

(1) Suppose $K = \mathbb{Q}(\sqrt{2})$ as in Example 1 above. Then the fixed field of $\text{Aut}(\mathbb{Q}(\sqrt{2})) = \text{Aut}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}) = \{1, \sigma\}$ will be the set of elements of $\mathbb{Q}(\sqrt{2})$ with

$$\sigma(a + b\sqrt{2}) = a + b\sqrt{2}$$

since everything is fixed by the identity automorphism. This is the equation

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

which is equivalent to $b = 0$, so the fixed field of $\text{Aut}(\mathbb{Q}(\sqrt{2})/\mathbb{Q})$ is just $\mathbb{Q}$.

(2) Suppose now that $K = \mathbb{Q}(\sqrt[3]{2})$ as in Example 2 above. In this case $\text{Aut}(K) = 1$, so that every element of K is fixed, i.e., the fixed field of $\text{Aut}(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q})$ is $\mathbb{Q}(\sqrt[3]{2})$.

Given a subfield F of K, the associated group is the collection of automorphisms of K which fix F. Given a group of automorphisms of K, the associated extension is defined by taking F to be the fixed field of the automorphisms. In the first example above, starting with the subfield $\mathbb{Q}$ of $\mathbb{Q}(\sqrt{2})$ one obtains the group $\{1, \sigma\}$ and starting with the group $\{1, \sigma\}$ one obtains the subfield $\mathbb{Q}$, so there is a "duality" between the two. In the second example, however, starting with the subfield $\mathbb{Q}$ of $\mathbb{Q}(\sqrt[3]{2})$ one obtains only the trivial group and starting with the trivial group one obtains the full field $\mathbb{Q}(\sqrt[3]{2})$.

An examination of the two examples suggests that for the second example there are "not enough" automorphisms to force the fixed field to be $\mathbb{Q}$ rather than the full $\mathbb{Q}(\sqrt[3]{2})$. This in turn seems to be due to the fact that the other roots of $x^3 - 2$, which are the only possible images of $\sqrt[3]{2}$ under an automorphism, are not elements of $\mathbb{Q}(\sqrt[3]{2})$. (Although even if they were we would need to check that the additional maps we could define were *automorphisms.*) We now make precise the notion of fields with "enough" automorphisms (leading to the definition of a *Galois* extension). As one might suspect even from these two examples (and we prove in the next section) these are related to splitting fields.

We first investigate the size of the automorphism group in the case of splitting fields.

Let F be a field and let E be the splitting field over F of $f(x) \in F[x]$. The main tool is Theorem 13.27 on the existence of extensions of isomorphisms, which states that any isomorphism $\varphi : F \xrightarrow{\sim} F'$ of F with F' can be extended to an isomorphism $\sigma : E \xrightarrow{\sim} E'$ between E and the splitting field E' for $f'(x) = \varphi(f(x)) \in F'[x]$.

We now show by induction on $[E : F]$ that the number of such extensions is at most $[E : F]$, with equality if $f(x)$ is separable over F. If $[E : F] = 1$ then $E = F$, $E' = F'$, $\sigma = \varphi$ and the number of extensions is 1. If $[E : F] > 1$ then $f(x)$ has at least one irreducible factor $p(x)$ of degree > 1 with corresponding irreducible factor $p'(x)$ of $f'(x)$. Let α be a fixed root of $p(x)$. If σ is any extension of φ to E, then σ restricted to the subfield $F(\alpha)$ of E is an isomorphism τ of $F(\alpha)$ with some subfield of E'. The isomorphism τ is completely determined by its action on α, i.e., by $\tau\alpha$, since α generates $F(\alpha)$ over F. Just as in Proposition 2, we see that $\tau\alpha$ must be some root β of $p'(x)$. Then we have a diagram

$$\begin{tikzcd}
\sigma : & E \arrow[r, "\sim"] \arrow[d, dash] & E' \arrow[d, dash] \\
\tau : & F(\alpha) \arrow[r, "\sim"] \arrow[d, dash] & F'(\beta) \arrow[d, dash] \\
\varphi : & F \arrow[r, "\sim"] & F'
\end{tikzcd}$$

Conversely, for any β a root of $p'(x)$ there are extensions τ and σ giving such a diagram (this is Theorem 13.8 and Theorem 13.27). Hence to count the number of extensions σ we need only count the possible number of these diagrams.

The number of extensions of φ to an isomorphism τ is equal to the number of distinct roots β of $p'(x)$. Since the degree of $p(x)$ and $p'(x)$ are both equal to $[F(\alpha) : F]$, we see that the number of extensions of φ to a τ is at most $[F(\alpha) : F]$, with equality if the roots of $p(x)$ are distinct.

Since E is also the splitting field of $f(x)$ over $F(\alpha)$, E' is the splitting field of $f'(x)$

over $F'(\beta)$, and $[E : F(\alpha)] < [E : F]$, we may apply our induction hypothesis to these field extensions. By induction, the number of extensions of τ to σ is $\leq [E : F(\alpha)]$, with equality if $f(x)$ has distinct roots.

From $[E : F] = [E : F(\alpha)][F(\alpha) : F]$ it follows that the number of extensions of φ to σ is $\leq [E : F]$. We have equality if $p(x)$ and $f(x)$ have distinct roots, which is equivalent to $f(x)$ having distinct roots since $p(x)$ is a factor of $f(x)$, completing the proof by induction.

In the particular case when $F = F'$ and φ is the identity map we have $f(x) = f'(x)$ and $E = E'$ so the isomorphisms of E to E' restricting to φ on F are the automorphisms of E fixing F. We state this as follows:

Proposition 5. Let E be the splitting field over F of the polynomial $f(x) \in F[x]$. Then

$$|\text{Aut}(E/F)| \leq [E : F]$$

with equality if $f(x)$ is separable over F.

Remark: While we were primarily interested in counting the automorphisms of E which fix F (which is the situation of $F = F'$, $\varphi = 1$ above), it would still have been necessary to consider the situation of more general φ (and different fields F') because of the induction step in the proof (which involves the fields $F(\alpha)$ and $F(\beta)$ for two roots of the same polynomial $p(x)$).

One can modify the proof above to show more generally that $|\text{Aut}(K/F)| \leq [K : F]$ for *any* finite extension K/F (we shall prove this in the next section from a slightly different point of view). This gives us a notion of field extensions with "enough" automorphisms.

Definition. Let K/F be a finite extension. Then K is said to be *Galois* over F and K/F is a *Galois* extension if $|\text{Aut}(K/F)| = [K : F]$. If K/F is Galois the group of automorphisms $\text{Aut}(K/F)$ is called the *Galois group of* K/F, denoted $\text{Gal}(K/F)$.

Remark: The Galois group of an extension K/F is sometimes defined to be the group of automorphisms $\text{Aut}(K/F)$ for all K/F. We have chosen the definition above so that the notation $\text{Gal}(K/F)$ will emphasize that the extension K/F has the maximal number of automorphisms.

Corollary 6. If K is the splitting field over F of a separable polynomial $f(x)$ then K/F is Galois.

We shall see in the next section that the converse is also true, which will completely characterize Galois extensions.

Note also that Corollary 6 implies that the splitting field of *any* polynomial over $\mathbb{Q}$ is Galois, since the splitting field of $f(x)$ is clearly the same as the splitting field of the product of the irreducible factors of $f(x)$ (i.e., the polynomial obtained by removing multiple factors), which is separable (Corollary 13.34).

Definition. If $f(x)$ is a separable polynomial over F, then the *Galois group of $f(x)$ over F* is the Galois group of the splitting field of $f(x)$ over F.

Examples

(1) The extension $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ is Galois with Galois group $\text{Gal}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}) = \{1, \sigma\} \cong \mathbb{Z}/2\mathbb{Z}$ where σ is the automorphism

$$\begin{aligned}\sigma : \mathbb{Q}(\sqrt{2}) &\xrightarrow{\sim} \mathbb{Q}(\sqrt{2})\\ a + b\sqrt{2} &\longmapsto a - b\sqrt{2}.\end{aligned}$$

(2) More generally, any quadratic extension K of any field F of characteristic different from 2 is Galois. This follows from the discussion of quadratic extensions following Corollary 13.13, which shows that any extension K of degree 2 of F (where the characteristic of F is not 2) is of the form $F(\sqrt{D})$ for some D hence is the splitting field of $x^2 - D$ (since if $\sqrt{D} \in K$ then also $-\sqrt{D} \in K$).

(3) The extension $\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}$ is not Galois since its group of automorphisms is only of order 1.

(4) The extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ is Galois over $\mathbb{Q}$ since it is the splitting field of the polynomial $(x^2 - 2)(x^2 - 3)$. Any automorphism σ is completely determined by its action on the generators $\sqrt{2}$ and $\sqrt{3}$, which must be mapped to $\pm\sqrt{2}$ and $\pm\sqrt{3}$, respectively. Hence the only possibilities for automorphisms are the maps

$$\begin{cases}\sqrt{2} \mapsto \sqrt{2}\\ \sqrt{3} \mapsto \sqrt{3}\end{cases} \quad \begin{cases}\sqrt{2} \mapsto -\sqrt{2}\\ \sqrt{3} \mapsto \sqrt{3}\end{cases} \quad \begin{cases}\sqrt{2} \mapsto \sqrt{2}\\ \sqrt{3} \mapsto -\sqrt{3}\end{cases} \quad \begin{cases}\sqrt{2} \mapsto -\sqrt{2}\\ \sqrt{3} \mapsto -\sqrt{3}\end{cases}.$$

Since the Galois group is of order 4, *all* these elements are in fact automorphisms of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$.

Define the automorphisms σ and τ by

$$\sigma : \begin{cases}\sqrt{2} \mapsto -\sqrt{2}\\ \sqrt{3} \mapsto \sqrt{3}\end{cases} \qquad \tau : \begin{cases}\sqrt{2} \mapsto \sqrt{2}\\ \sqrt{3} \mapsto -\sqrt{3}\end{cases}$$

or, more explicitly, by

$$\begin{aligned}\sigma &: a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mapsto a - b\sqrt{2} + c\sqrt{3} - d\sqrt{6}\\ \tau &: a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mapsto a + b\sqrt{2} - c\sqrt{3} - d\sqrt{6}\end{aligned}$$

(since, for example,

$$\sigma(\sqrt{6}) = \sigma(\sqrt{2}\sqrt{3}) = \sigma(\sqrt{2})\sigma(\sqrt{3}) = (-\sqrt{2})(\sqrt{3}) = -\sqrt{6}\quad).$$

Then $\sigma^2(\sqrt{2}) = \sigma(\sigma\sqrt{2}) = \sigma(-\sqrt{2}) = \sqrt{2}$ and clearly $\sigma^2(\sqrt{3}) = \sqrt{3}$. Hence $\sigma^2 = 1$ is the identity automorphism. Similarly, $\tau^2 = 1$. The automorphism $\sigma\tau$ can be easily computed:

$$\sigma\tau(\sqrt{2}) = \sigma(\tau(\sqrt{2})) = \sigma(\sqrt{2}) = -\sqrt{2}$$

and

$$\sigma\tau(\sqrt{3}) = \sigma(\tau(\sqrt{3})) = \sigma(-\sqrt{3}) = -\sqrt{3}$$

so that $\sigma\tau$ is the remaining nontrivial automorphism in the Galois group. Since this automorphism also evidently has order 2 in the Galois group, we have

$$\text{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}) = \{1, \sigma, \tau, \sigma\tau\}$$

i.e., the Galois group is isomorphic to the Klein 4-group.

Associated to each subgroup of $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q})$ is the corresponding fixed subfield of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$. For example, the subfield corresponding to $\{1, \sigma\tau\}$ is the set of elements fixed by the map

$$\sigma\tau : a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mapsto a - b\sqrt{2} - c\sqrt{3} + d\sqrt{6}$$

which is the set of elements $a+d\sqrt{6}$, i.e., the field $\mathbb{Q}(\sqrt{6})$. One can similarly determine the fixed fields for the other subgroups of the Galois group:

subgroup	fixed field
$\{1\}$	$\mathbb{Q}(\sqrt{2}, \sqrt{3})$
$\{1, \sigma\}$	$\mathbb{Q}(\sqrt{3})$
$\{1, \sigma\tau\}$	$\mathbb{Q}(\sqrt{6})$
$\{1, \tau\}$	$\mathbb{Q}(\sqrt{2})$
$\{1, \sigma, \tau, \sigma\tau\}$	$\mathbb{Q}$

(5) The splitting field of $x^3 - 2$ over $\mathbb{Q}$ is Galois of degree 6. The roots of this equation are $\sqrt[3]{2}$, $\rho\sqrt[3]{2}$, $\rho^2\sqrt[3]{2}$ where $\rho = \zeta_3 = \dfrac{-1+\sqrt{-3}}{2}$ is a primitive cube root of unity. Hence the splitting field can be written $\mathbb{Q}(\sqrt[3]{2}, \rho\sqrt[3]{2})$. Any automorphism maps each of these two elements to one of the roots of $x^3 - 2$, giving 9 possibilities, but since the Galois group has order 6 not every such map is an automorphism of the field.

To determine the Galois group we use a more convenient set of generators, namely $\sqrt[3]{2}$ and ρ. Then any automorphism σ maps $\sqrt[3]{2}$ to one of $\sqrt[3]{2}$, $\rho\sqrt[3]{2}$, $\rho^2\sqrt[3]{2}$ and maps ρ to ρ or $\rho^2 = \dfrac{-1-\sqrt{-3}}{2}$ since these are the roots of the cyclotomic polynomial $\Phi_3(x) = x^2 + x + 1$. Since σ is completely determined by its action on these two elements this gives only 6 possibilities and so each of these possibilities is actually an automorphism. To give these automorphisms explicitly, let σ and τ be the automorphisms defined by

$$\sigma : \begin{cases} \sqrt[3]{2} \mapsto \rho\sqrt[3]{2} \\ \rho \mapsto \rho \end{cases} \qquad \tau : \begin{cases} \sqrt[3]{2} \mapsto \sqrt[3]{2} \\ \rho \mapsto \rho^2 = -1 - \rho. \end{cases}$$

As before, these can be given explicitly on the elements of $\mathbb{Q}(\sqrt[3]{2}, \rho)$, which are linear combinations of the basis $\{1, \sqrt[3]{2}, (\sqrt[3]{2})^2, \rho, \rho\sqrt[3]{2}, \rho(\sqrt[3]{2})^2\}$. For example

$$\begin{aligned} \sigma(\rho\sqrt[3]{2}) &= (\rho)(\rho\sqrt[3]{2}) = \rho^2\sqrt[3]{2} = (-1-\rho)\sqrt[3]{2} \\ &= -\sqrt[3]{2} - \rho\sqrt[3]{2} \end{aligned}$$

and we may similarly determine the action of σ on the other basis elements. This gives

$$\begin{aligned} \sigma : \quad & a + b\sqrt[3]{2} + c\sqrt[3]{4} + d\rho + e\rho\sqrt[3]{2} + f\rho\sqrt[3]{4} \quad \longmapsto \\ & a - e\sqrt[3]{2} + (f-c)\sqrt[3]{4} + d\rho + (b-e)\rho\sqrt[3]{2} - c\rho\sqrt[3]{4}. \end{aligned} \tag{14.1}$$

The other elements of the Galois group are

$$1 : \begin{cases} \sqrt[3]{2} \mapsto \sqrt[3]{2} \\ \rho \mapsto \rho \end{cases} \qquad \sigma^2 : \begin{cases} \sqrt[3]{2} \mapsto \rho^2\sqrt[3]{2} \\ \rho \mapsto \rho \end{cases}$$

Computing $\sigma\tau$ we have

$$\sigma\tau : \begin{cases} \sqrt[3]{2} \overset{\tau}{\mapsto} \sqrt[3]{2} \overset{\sigma}{\mapsto} \rho\sqrt[3]{2} \\ \rho \overset{\tau}{\mapsto} \rho^2 \overset{\sigma}{\mapsto} \rho^2 \end{cases}$$

i.e.,

$$\sigma\tau : \begin{cases} \sqrt[3]{2} \mapsto \rho\sqrt[3]{2} \\ \rho \mapsto \rho^2 \end{cases}$$

so that $\sigma\tau = \tau\sigma^2$. Similarly one computes that $\sigma^3 = \tau^2 = 1$. Hence

$$\mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2}, \zeta_3)/\mathbb{Q}) = \langle \sigma, \tau \rangle \cong S_3$$

is the symmetric group on 3 letters. Alternatively (and less computationally), since $G = \mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2}, \zeta_3)/\mathbb{Q})$ acts as permutations of the 3 roots of $x^3 - 2$, G is a subgroup of S_3, hence must be S_3 since it is of order 6. The computations above explicitly identify the automorphisms in G and give an explicit isomorphism of G with S_3.

As in the previous example we can determine the fixed fields for any of the subgroups of the Galois group. For example, consider the fixed field of the subgroup $\{1, \sigma, \sigma^2\}$ generated by σ. These are just the elements fixed by σ (given explicitly in equation (1)) since if an element is fixed by σ then it is also fixed by σ^2. (In general, the fixed field of some subgroup is the field fixed by a set of generators for the subgroup.) The elements fixed by σ are those with

$$a = a \quad b = -e \quad c = f - c \quad d = d \quad e = b - e \quad f = -c$$

which is equivalent to $b = c = f = e = 0$. Hence the fixed field of $\{1, \sigma, \sigma^2\}$ is the field $\mathbb{Q}(\rho)$.

Remark: This example shows that some care must be exercised in determining Galois groups from the actions on generators. As mentioned, not every map taking $\sqrt[3]{2}$ and $\rho\sqrt[3]{2}$ to roots of $x^3 - 2$ gives rise to an automorphism of the field (for example, the map

$$\begin{aligned} \sqrt[3]{2} &\mapsto \rho\sqrt[3]{2} \\ \rho\sqrt[3]{2} &\mapsto \rho\sqrt[3]{2} \end{aligned}$$

clearly cannot be an automorphism since it is evidently not an injection). The point is that there may be (sometimes very subtle) algebraic relations among the generators and these relations must be respected by an automorphism. For example, the quotient of the generators here is ρ, which is mapped to 1 and not to a root of the minimal polynomial for ρ. Put another way, the quotient of these generators satisfies a quadratic equation and this map does not respect that property.

For another (less trivial) example, compare with the discussion of the splitting field of $x^8 - 2$ in Section 2.

(6) As in Example 3, the field $\mathbb{Q}(\sqrt[4]{2})$ is not Galois over $\mathbb{Q}$ since any automorphism is determined by where it sends $\sqrt[4]{2}$ and of the four possibilities $\{\pm\sqrt[4]{2}, \pm i\sqrt[4]{2}\}$, only two are elements of the field (the two real roots).

Note that we have

$$\overbrace{\underbrace{\mathbb{Q} \quad \subset \quad \mathbb{Q}(\sqrt{2})}_{2} \quad \underbrace{\subset \quad \mathbb{Q}(\sqrt[4]{2})}_{2}}^{4}$$

where $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ and $\mathbb{Q}(\sqrt[4]{2})/\mathbb{Q}(\sqrt{2})$ are both Galois extensions by Example 2 since both are quadratic extensions. This shows that a Galois extension of a Galois extension is not necessarily Galois.

(7) The extension of finite fields $\mathbb{F}_{p^n}/\mathbb{F}_p$ constructed after Proposition 13.37 is Galois by Corollary 6 since $\mathbb{F}_{p^n}$ is the splitting field over $\mathbb{F}_p$ of the separable polynomial $x^{p^n} - x$. It follows that the group of automorphisms for this extension is of order n. The injective homomorphism

$$\begin{aligned} \sigma : \mathbb{F}_{p^n} &\to \mathbb{F}_{p^n} \\ \alpha &\mapsto \alpha^p \end{aligned}$$

of Proposition 13.35 is surjective in this case since $\mathbb{F}_{p^n}$ is finite, hence is an isomorphism. This gives an automorphism of $\mathbb{F}_{p^n}$, called the *Frobenius* automorphism, which we shall denote by σ_p. Iterating σ_p we have $\sigma_p^2(\alpha) = \sigma_p(\sigma_p(\alpha)) = (\alpha^p)^p = \alpha^{p^2}$. Similarly we have

$$\sigma_p^i(\alpha) = \alpha^{p^i} \qquad i = 0, 1, 2, \ldots$$

Since $\alpha^{p^n} = \alpha$, we see that $\sigma_p^{p^n} = 1$ is the identity automorphism. No lower power of σ_p can be the identity, since this would imply $\alpha^{p^i} = \alpha$ for all $\alpha \in \mathbb{F}_{p^n}$ for some $i < n$, which is impossible since there are only p^i roots of this equation. It follows that σ_p is of order n in the Galois group, which means that $\mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$ is *cyclic* of order n, with the Frobenius automorphism σ_p as generator.

(8) The inseparable extension $\mathbb{F}_2(x)$ over $\mathbb{F}_2(t)$ where $x^2 - t = 0$ considered in Section 13.5 is not Galois. Any automorphism of this degree 2 extension is determined by its action on x, which must be sent to a root of the equation $x^2 - t$. We have already seen that there is only one root of this equation (with multiplicity 2) since we are in a field of characteristic 2. Hence the extension has only the trivial automorphism. Note that $\mathbb{F}_2(x)$ is the splitting field for $x^2 - t$ over $\mathbb{F}_2(t)$, so this example shows the separability condition in Corollary 6 is necessary.

EXERCISES

1. **(a)** Show that if the field K is generated over F by the elements $\alpha_1, \ldots, \alpha_n$ then an automorphism σ of K fixing F is uniquely determined by $\sigma(\alpha_1), \ldots, \sigma(\alpha_n)$. In particular show that an automorphism fixes K if and only if it fixes a set of generators for K.

(b) Let $G \leq \mathrm{Gal}(K/F)$ be a subgroup of the Galois group of the extension K/F and suppose $\sigma_1, \ldots, \sigma_k$ are generators for G. Show that the subfield E/F is fixed by G if and only if it is fixed by the generators $\sigma_1, \ldots, \sigma_k$.

2. Let τ be the map $\tau : \mathbb{C} \to \mathbb{C}$ defined by $\tau(a+bi) = a - bi$ (*complex conjugation*). Prove that τ is an automorphism of $\mathbb{C}$.

3. Determine the fixed field of complex conjugation on $\mathbb{C}$.

4. Prove that $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{3})$ are not isomorphic.

5. Determine the automorphisms of the extension $\mathbb{Q}(\sqrt[4]{2})/\mathbb{Q}(\sqrt{2})$ explicitly.

6. Let k be a field.
   (a) Show that the mapping $\varphi : k[t] \to k[t]$ defined by $\varphi(f(t)) = f(at+b)$ for fixed $a, b \in k$, $a \neq 0$ is an automorphism of $k[t]$ which is the identity on k.
   (b) Conversely, let φ be an automorphism of $k[t]$ which is the identity on k. Prove that there exist $a, b \in k$ with $a \neq 0$ such that $\varphi(f(t)) = f(at+b)$ as in (a).

7. This exercise determines $\mathrm{Aut}(\mathbb{R}/\mathbb{Q})$.
   (a) Prove that any $\sigma \in \mathrm{Aut}(\mathbb{R}/\mathbb{Q})$ takes squares to squares and takes positive reals to positive reals. Conclude that $a < b$ implies $\sigma a < \sigma b$ for every $a, b \in \mathbb{R}$.
   (b) Prove that $-\frac{1}{m} < a - b < \frac{1}{m}$ implies $-\frac{1}{m} < \sigma a - \sigma b < \frac{1}{m}$ for every positive integer m. Conclude that σ is a continuous map on $\mathbb{R}$.
   (c) Prove that any continuous map on $\mathbb{R}$ which is the identity on $\mathbb{Q}$ is the identity map, hence $\mathrm{Aut}(\mathbb{R}/\mathbb{Q}) = 1$.

8. Prove that the automorphisms of the rational function field $k(t)$ which fix k are precisely the *fractional linear transformations* determined by $t \mapsto \dfrac{at+b}{ct+d}$ for $a, b, c, d \in k, ad-bc \neq 0$ (so $f(t) \in k(t)$ maps to $f(\dfrac{at+b}{ct+d})$) (cf. Exercise 18 of Section 13.2).

9. Determine the fixed field of the automorphism $t \mapsto t + 1$ of $k(t)$.

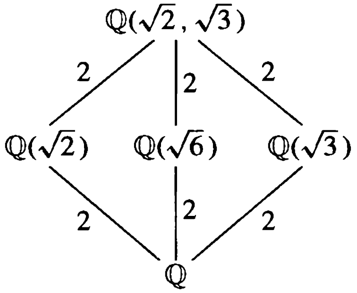
10. Let K be an extension of the field F. Let $\varphi : K \to K'$ be an isomorphism of K with a field K' which maps F to the subfield F' of K'. Prove that the map $\sigma \mapsto \varphi\sigma\varphi^{-1}$ defines a group isomorphism $\mathrm{Aut}(K/F) \xrightarrow{\sim} \mathrm{Aut}(K'/F')$.

14.2 THE FUNDAMENTAL THEOREM OF GALOIS THEORY

In the Galois extension $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q})$ considered in the previous section, there was a strong similarity between the diagram of subgroups of the Galois group:

$$\begin{tikzcd}
& \{1\} \arrow[dl, dash, "2"'] \arrow[d, dash, "2"] \arrow[dr, dash, "2"] & \\
\{1, \tau\} \arrow[dr, dash, "2"'] & \{1, \sigma\tau\} \arrow[d, dash, "2"] & \{1, \sigma\} \arrow[dl, dash, "2"] \\
& \{1, \sigma, \tau, \sigma\tau\} &
\end{tikzcd}$$

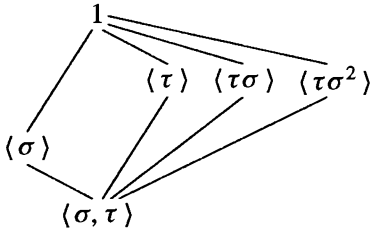
and the diagram of corresponding fixed fields



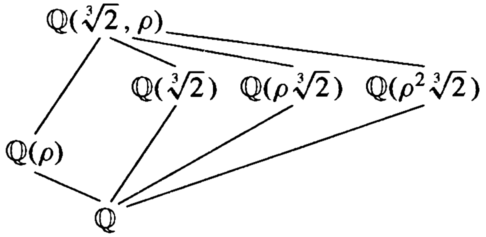
(we have inverted the lattice of subgroups because of the inclusion-reversing nature of the correspondence).

Note that this is also the diagram of *all* known subfields of the extension and that in this case each of the subfields is also a Galois extension of $\mathbb{Q}$.

In a similar way there is a strong similarity between the diagram


1
$\langle \tau \rangle$
$\langle \tau\sigma \rangle$
$\langle \tau\sigma^2 \rangle$
$\langle \sigma \rangle$
$\langle \sigma, \tau \rangle$

of subgroups of the Galois group and the diagram of known subfields for the splitting field of $x^3 - 2$:


$\mathbb{Q}(\sqrt[3]{2},\rho)$
$\mathbb{Q}(\sqrt[3]{2})$
$\mathbb{Q}(\rho\sqrt[3]{2})$
$\mathbb{Q}(\rho^2\sqrt[3]{2})$
$\mathbb{Q}(\rho)$
$\mathbb{Q}$

where the subfields in the second diagram are precisely the fixed fields of the subgroups in the first diagram.

Note in this pair of diagrams only the subgroup $\langle \sigma \rangle$ generated by σ is normal in S_3 and that the subfield $\mathbb{Q}(\rho)$ is the only subfield Galois over $\mathbb{Q}$.

The Fundamental Theorem of Galois Theory states that the relations observed in the two examples above are not coincidental and hold for any Galois extension. Before proving this we first develop some preliminary results on *group characters*, of which field automorphisms give particular examples.

Definition. A *character*[1] χ of a group G with values in a field L is a homomorphism from G to the multiplicative group of L:

$$\chi : G \to L^\times$$

i.e., $\chi(g_1g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in G$ and $\chi(g)$ is a nonzero element of L for all $g \in G$.

Definition. The characters $\chi_1, \chi_2, \dots, \chi_n$ of G are said to be *linearly independent* over L if they are linearly independent as functions on G, i.e., if there is no nontrivial relation

$$a_1\chi_1 + a_2\chi_2 + \cdots + a_n\chi_n = 0 \qquad (a_1, \dots, a_n \in L \text{ not all } 0) \tag{14.2}$$

as a function on G (that is, $a_1\chi_1(g) + a_2\chi_2(g) + \cdots + a_n\chi_n(g) = 0$ for all $g \in G$).

Theorem 7. *(Linear Independence of Characters)* If $\chi_1, \chi_2, \dots, \chi_n$ are distinct characters of G with values in L then they are linearly independent over L.

Proof: Suppose the characters were linearly dependent. Among all the linear dependence relations (2) above, choose one with the minimal number m of nonzero coefficients a_i. We may suppose (by renumbering, if necessary) that the m nonzero coefficients are $a_1, a_2, \dots, a_m$:

$$a_1\chi_1 + a_2\chi_2 + \cdots + a_m\chi_m = 0.$$

Then for any $g \in G$ we have

$$a_1\chi_1(g) + a_2\chi_2(g) + \cdots + a_m\chi_m(g) = 0. \tag{14.3}$$

Let g_0 be an element with $\chi_1(g_0) \neq \chi_m(g_0)$ (which exists, since $\chi_1 \neq \chi_m$). Since (3) holds for every element of G, in particular we have

$$a_1\chi_1(g_0g) + a_2\chi_2(g_0g) + \cdots + a_m\chi_m(g_0g) = 0$$

i.e.,

$$a_1\chi_1(g_0)\chi_1(g) + a_2\chi_2(g_0)\chi_2(g) + \cdots + a_m\chi_m(g_0)\chi_m(g) = 0. \tag{14.4}$$

Multiplying equation (3) by $\chi_m(g_0)$ and subtracting from equation (4) we obtain

$$\begin{aligned}[\chi_m(g_0) - \chi_1(g_0)]a_1\chi_1(g) + [\chi_m(g_0) - \chi_2(g_0)]a_2\chi_2(g) + \cdots\\ + [\chi_m(g_0) - \chi_{m-1}(g_0)]a_{m-1}\chi_{m-1}(g) = 0,\end{aligned}$$

which holds for all $g \in G$. But the first coefficient is nonzero and this is a relation with fewer nonzero coefficients, a contradiction.

Consider now an injective homomorphism σ of a field K into a field L, called an *embedding* of K into L. Then in particular σ is a homomorphism of the multiplicative group $G = K^\times$ into the multiplicative group $L^\times$, so σ may be viewed as a character of $K^\times$ with values in L. Note also that this character contains all of the useful information about the values of σ viewed simply as a *function* on K, since the only point of K not considered in $K^\times$ is 0, and we know σ maps 0 to 0.

[1] This is the definition of a *linear* character. More general characters will be studied in Chapter 18.

Corollary 8. If $\sigma_1, \sigma_2, \ldots, \sigma_n$ are distinct embeddings of a field K into a field L, then they are linearly independent as functions on K. In particular distinct automorphisms of a field K are linearly independent as functions on K.

We now use Corollary 8 to prove the fundamental relation between the orders of subgroups of the automorphism group of a field K and the degrees of the extensions over their fixed fields.

Theorem 9. Let $G = \{\sigma_1 = 1, \sigma_2, \ldots, \sigma_n\}$ be a subgroup of automorphisms of a field K and let F be the fixed field. Then

$$[K : F] = n = |G|.$$

Proof: Suppose first that $n > [K : F]$ and let $\omega_1, \omega_2, \ldots, \omega_m$ be a basis for K over F ($m = [K : F]$). Then the system

$$\begin{gathered}\sigma_1(\omega_1)x_1 + \sigma_2(\omega_1)x_2 + \cdots + \sigma_n(\omega_1)x_n = 0\\ \vdots\\ \sigma_1(\omega_m)x_1 + \sigma_2(\omega_m)x_2 + \cdots + \sigma_n(\omega_m)x_n = 0\end{gathered}$$

of m equations in n unknowns $x_1, x_2, \ldots, x_n$ has a nontrivial solution $\beta_1, \beta_2, \ldots, \beta_n$ in K since by assumption there are more unknowns than equations.

Let $a_1, a_2, \ldots, a_m$ be m arbitrary elements of F. The field F is by definition fixed by $\sigma_1, \ldots, \sigma_n$ so each of these elements is fixed by every σ_i, i.e., $\sigma_i(a_j) = a_j$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Multiplying the first equation above by a_1, the second by $a_2, \ldots$, the last by a_m then gives the system of equations

$$\begin{gathered}\sigma_1(a_1\omega_1)\beta_1 + \sigma_2(a_1\omega_1)\beta_2 + \cdots + \sigma_n(a_1\omega_1)\beta_n = 0\\ \vdots\\ \sigma_1(a_m\omega_m)\beta_1 + \sigma_2(a_m\omega_m)\beta_2 + \cdots + \sigma_n(a_m\omega_m)\beta_n = 0.\end{gathered}$$

Adding these equations we see that there are elements $\beta_1, \ldots, \beta_n$ in K, not all 0, satisfying

$$\sigma_1(a_1\omega_1 + a_2\omega_2 + \cdots + a_m\omega_m)\beta_1 + \cdots + \sigma_n(a_1\omega_1 + a_2\omega_2 + \cdots + a_m\omega_m)\beta_n = 0$$

for all choices of $a_1, \ldots, a_m$ in F. Since $\omega_1, \ldots, \omega_m$ is an F-basis for K, every $\alpha \in K$ is of the form $a_1\omega_1 + a_2\omega_2 + \cdots + a_m\omega_m$, so the previous equation means

$$\sigma_1(\alpha)\beta_1 + \cdots + \sigma_n(\alpha)\beta_n = 0$$

for all $\alpha \in K$. But this means the distinct automorphisms $\sigma_1, \ldots, \sigma_n$ are linearly dependent over K, contradicting Corollary 8.

We have proved $n \leq [K : F]$. Note that we have so far not used the fact that $\sigma_1, \sigma_2, \ldots, \sigma_n$ are the elements of a *group*.

Suppose now that $n < [K : F]$. Then there are more than n F-linearly independent elements of K, say $\alpha_1, \ldots, \alpha_{n+1}$. The system

$$\begin{gathered}\sigma_1(\alpha_1)x_1 + \sigma_1(\alpha_2)x_2 + \cdots + \sigma_1(\alpha_{n+1})x_{n+1} = 0\\ \vdots\\ \sigma_n(\alpha_1)x_1 + \sigma_n(\alpha_2)x_2 + \cdots + \sigma_n(\alpha_{n+1})x_{n+1} = 0\end{gathered} \tag{14.5}$$

of n equations in $n+1$ unknowns $x_1, \dots, x_{n+1}$ has a solution $\beta_1, \dots, \beta_{n+1}$ in K where not all the $\beta_i, i = 1, 2, \dots, n+1$ are 0. If all the elements of the solution $\beta_1, \dots, \beta_{n+1}$ were elements of F then the first equation (recall $\sigma_1 = 1$ is the identity automorphism) would contradict the linear independence over F of $\alpha_1, \alpha_2, \dots, \alpha_{n+1}$. Hence at least one $\beta_i, i = 1, 2, \dots, n+1$, is not an element of F.

Among all the nontrivial solutions $(\beta_1, \dots, \beta_{n+1})$ of the system (5) choose one with the minimal number r of nonzero β_i. By renumbering if necessary we may assume $\beta_1, \dots, \beta_r$ are nonzero. Dividing the equations by β_r we may also assume $\beta_r = 1$. We have already seen that at least one of $\beta_1, \dots, \beta_{r-1}, 1$ is not an element of F (which shows in particular that $r > 1$), say $\beta_1 \notin F$. Then our system of equations reads

$$\begin{array}{c} \sigma_1(\alpha_1)\beta_1 + \cdots + \sigma_1(\alpha_{r-1})\beta_{r-1} + \sigma_1(\alpha_r) = 0 \\ \vdots \\ \sigma_n(\alpha_1)\beta_1 + \cdots + \sigma_n(\alpha_{r-1})\beta_{r-1} + \sigma_n(\alpha_r) = 0 \end{array} \tag{14.6}$$

or more briefly

$$\sigma_i(\alpha_1)\beta_1 + \cdots + \sigma_i(\alpha_{r-1})\beta_{r-1} + \sigma_i(\alpha_r) = 0 \qquad i = 1, 2, \dots, n. \tag{14.7}$$

Since $\beta_1 \notin F$, there is an automorphism σ_{k_0} ($k_0 \in \{1, 2, \dots, n\}$) with $\sigma_{k_0}\beta_1 \neq \beta_1$. If we apply the automorphism σ_{k_0} to the equations in (6), we obtain the system of equations

$$\sigma_{k_0}\sigma_j(\alpha_1)\sigma_{k_0}(\beta_1) + \cdots + \sigma_{k_0}\sigma_j(\alpha_{r-1})\sigma_{k_0}(\beta_{r-1}) + \sigma_{k_0}\sigma_j(\alpha_r) = 0 \tag{14.8}$$

for $j = 1, 2, \dots, n$. But the elements

$$\sigma_{k_0}\sigma_1, \ \sigma_{k_0}\sigma_2, \ \dots, \ \sigma_{k_0}\sigma_n$$

are the same as the elements

$$\sigma_1, \ \sigma_2, \ \dots, \ \sigma_n$$

in some order since these elements form a *group*. In other words, if we define the index i by $\sigma_{k_0}\sigma_j = \sigma_i$ then i and j both run over the set $\{1, 2, \dots, n\}$. Hence the equations in (8) can be written

$$\sigma_i(\alpha_1)\sigma_{k_0}(\beta_1) + \cdots + \sigma_i(\alpha_{r-1})\sigma_{k_0}(\beta_{r-1}) + \sigma_i(\alpha_r) = 0. \tag{14.8$'$}$$

If we now subtract the equations in (8′) from those in (7) we obtain the system

$$\sigma_i(\alpha_1)[\beta_1 - \sigma_{k_0}(\beta_1)] + \cdots + \sigma_i(\alpha_{r-1})[\beta_{r-1} - \sigma_{k_0}(\beta_{r-1})] = 0$$

for $i = 1, 2, \dots, n$. But this is a solution to the system of equations (5) with

$$x_1 = \beta_1 - \sigma_{k_0}(\beta_1) \neq 0$$

(by the choice of k_0), hence is nontrivial and has fewer than r nonzero x_i. This is a contradiction and completes the proof.

Our first use of this result is to prove that the inequality of Proposition 5 holds for any finite extension K/F.

Corollary 10. Let K/F be any finite extension. Then

$$|\mathrm{Aut}(K/F)| \leq [K : F]$$

with equality if and only if F is the fixed field of $\mathrm{Aut}(K/F)$. Put another way, K/F is Galois if and only if F is the fixed field of $\mathrm{Aut}(K/F)$.

Proof: Let F_1 be the fixed field of $\mathrm{Aut}(K/F)$, so that

$$F \subseteq F_1 \subseteq K.$$

By Theorem 9, $[K : F_1] = |\mathrm{Aut}(K/F)|$. Hence $[K : F] = |\mathrm{Aut}(K/F)|[F_1 : F]$, which proves the corollary.

Corollary 11. Let G be a finite subgroup of automorphisms of a field K and let F be the fixed field. Then every automorphism of K fixing F is contained in G, i.e., $\mathrm{Aut}(K/F) = G$, so that K/F is Galois, with Galois group G.

Proof: By definition F is fixed by all the elements of G so we have $G \leq \mathrm{Aut}(K/F)$ (and the question is whether there are any automorphisms of K fixing F not in G i.e., whether this containment is proper). Hence $|G| \leq |\mathrm{Aut}(K/F)|$. By the theorem we have $|G| = [K : F]$ and by the previous corollary $|\mathrm{Aut}(K/F)| \leq [K : F]$. This gives

$$[K : F] = |G| \leq |\mathrm{Aut}(K/F)| \leq [K : F]$$

and it follows that we must have equalities throughout, proving the corollary.

Corollary 12. If $G_1 \neq G_2$ are distinct finite subgroups of automorphisms of a field K then their fixed fields are also distinct.

Proof: Suppose F_1 is the fixed field of G_1 and F_2 is the fixed field of G_2. If $F_1 = F_2$ then by definition F_1 is fixed by G_2. By the previous corollary any automorphism fixing F_1 is contained in G_1, hence $G_2 \leq G_1$. Similarly $G_1 \leq G_2$ and so $G_1 = G_2$.

By the corollaries above we see that taking the fixed fields for distinct finite subgroups of $\mathrm{Aut}(K)$ gives distinct subfields of K over which K is Galois. Further, the degrees of the extensions are given by the orders of the subgroups. We saw this explicitly for the fields $K = \mathbb{Q}(\sqrt{2}, \sqrt{3})$ and $K = \mathbb{Q}(\sqrt[3]{2}, \rho)$ above. A portion of the Fundamental Theorem states that these are *all* the subfields of K.

The next result provides the converse of Proposition 5 and characterizes Galois extensions.

Theorem 13. The extension K/F is Galois if and only if K is the splitting field of some separable polynomial over F. Furthermore, if this is the case then every irreducible polynomial with coefficients in F which has a root in K is separable and has all its roots in K (so in particular K/F is a separable extension).

Proof: Proposition 5 proves that the splitting field of a separable polynomial is Galois.

We now show that if K/F is Galois then every irreducible polynomial $p(x)$ in $F[x]$ having a root in K splits completely in K. Set $G = \text{Gal}(K/F)$. Let $\alpha \in K$ be a root of $p(x)$ and consider the elements

$$\alpha, \sigma_2(\alpha), \dots, \sigma_n(\alpha) \in K \tag{14.9}$$

where $\{1, \sigma_2, \dots, \sigma_n\}$ are the elements of $\text{Gal}(K/F)$. Let

$$\alpha, \alpha_2, \alpha_3, \dots, \alpha_r$$

denote the *distinct* elements in (9). If $\tau \in G$ then since G is a group the elements $\{\tau, \tau\sigma_2, \dots, \tau\sigma_n\}$ are the same as the elements $\{1, \sigma_2, \dots, \sigma_n\}$ in some order. It follows that applying $\tau \in G$ to the elements in (9) simply permutes them, so in particular applying τ to $\alpha, \alpha_2, \alpha_3, \dots, \alpha_r$ also permutes these elements. The polynomial

$$f(x) = (x - \alpha)(x - \alpha_2) \cdots (x - \alpha_r)$$

therefore has coefficients which are fixed by all the elements of G since the elements of G simply permute the factors. Hence the coefficients lie in the fixed field of G, which by Corollary 10 is the field F. Hence $f(x) \in F[x]$.

Since $p(x)$ is irreducible and has α as a root, $p(x)$ is the minimal polynomial for α over F, hence divides any polynomial with coefficients in F having α as a root (this is Proposition 13.9). It follows that $p(x)$ divides $f(x)$ in $F[x]$ and since $f(x)$ obviously divides $p(x)$ in $K[x]$ by Proposition 2, we have

$$p(x) = f(x).$$

In particular, this shows that $p(x)$ is separable and that all its roots lie in K (in fact they are among the elements $\alpha, \sigma_2\alpha, \dots, \sigma_n\alpha$), proving the last statement of the theorem.

To complete the proof, suppose K/F is Galois and let $\omega_1, \omega_2, \dots, \omega_n$ be a basis for K/F. Let $p_i(x)$ be the minimal polynomial for ω_i over F, $i = 1, 2, \dots, n$. Then by what we have just proved, $p_i(x)$ is separable and has all its roots in K. Let $g(x)$ be the polynomial obtained by removing any multiple factors in the product $p_1(x) \cdots p_n(x)$ (the "squarefree part"). Then the splitting field of the two polynomials is the same and this field is K (all the roots lie in K, so K contains the splitting field, but $\omega_1, \omega_2, \dots, \omega_n$ are among the roots, so the splitting field contains K). Hence K is the splitting field of the separable polynomial $g(x)$.

Definition. Let K/F be a Galois extension. If $\alpha \in K$ the elements $\sigma\alpha$ for σ in $\text{Gal}(K/F)$ are called the *conjugates* (or *Galois conjugates*) of α over F. If E is a subfield of K containing F, the field $\sigma(E)$ is called the *conjugate field* of E over F.

The proof of the theorem shows that in a Galois extension K/F the other roots of the minimal polynomial over F of any element $\alpha \in K$ are precisely the distinct conjugates of α under the Galois group of K/F.

The second statement in this theorem also shows that K is not Galois over F if we can find even one irreducible polynomial over F having a root in K but not having *all* its roots in K. This justifies in a very strong sense the intuition from earlier examples that Galois extensions are extensions with "enough" distinct roots of irreducible polynomials (namely, if it contains one root then it contains all the roots).

Finally, notice that we now have 4 characterizations of Galois extensions K/F:

(1) splitting fields of separable polynomials over F

(2) fields where F is precisely the set of elements fixed by $\mathrm{Aut}(K/F)$ (in general, the fixed field may be larger than F)

(3) fields with $[K : F] = |\mathrm{Aut}(K/F)|$ (the original definition)

(4) finite, normal and separable extensions.

Theorem 14. *(Fundamental Theorem of Galois Theory)* Let K/F be a Galois extension and set $G = \mathrm{Gal}(K/F)$. Then there is a bijection

$$\left\{\begin{array}{cc} & K \\ \text{subfields } E & | \\ \text{of } K & E \\ \text{containing } F & | \\ & F \end{array}\right\} \longleftrightarrow \left\{\begin{array}{cc} & 1 \\ \text{subgroups } H & | \\ \text{of } G & H \\ & | \\ & G \end{array}\right\}$$

given by the correspondences

$$\begin{array}{ccc} & E & \longrightarrow & \left\{\begin{array}{c}\text{the elements of } G \\ \text{fixing } E\end{array}\right\} \\ \left\{\begin{array}{c}\text{the fixed field} \\ \text{of } H\end{array}\right\} & \longleftarrow & H & \end{array}$$

which are inverse to each other. Under this correspondence,

(1) (inclusion reversing) If E_1, E_2 correspond to H_1, H_2, respectively, then $E_1 \subseteq E_2$ if and only if $H_2 \leq H_1$

(2) $[K : E] = |H|$ and $[E : F] = |G : H|$, the index of H in G:

$$\begin{array}{ccl} K & & \\ | & \} & |H| \\ E & & \\ | & \} & |G : H| \\ F & & \end{array}$$

(3) K/E is always Galois, with Galois group $\mathrm{Gal}(K/E) = H$:

$$\begin{array}{cl} K & \\ | & H \\ E & \end{array}$$

(4) E is Galois over F if and only if H is a normal subgroup in G. If this is the case, then the Galois group is isomorphic to the quotient group

$$\mathrm{Gal}(E/F) \cong G/H.$$

More generally, even if H is not necessarily normal in G, the isomorphisms of E (into a fixed algebraic closure of F containing K) which fix F are in one to one correspondence with the cosets $\{\sigma H\}$ of H in G.

(5) If E_1, E_2 correspond to H_1, H_2, respectively, then the intersection $E_1 \cap E_2$ corresponds to the group $\langle H_1, H_2 \rangle$ generated by H_1 and H_2 and the composite field E_1E_2 corresponds to the intersection $H_1 \cap H_2$. Hence the lattice of subfields of K containing F and the lattice of subgroups of G are "dual" (the lattice diagram for one is the lattice diagram for the other turned upside down).

Proof: Given any subgroup H of G we obtain a unique fixed field $E = K_H$ by Corollary 12. This shows that the correspondence above is injective from right to left.

If K is the splitting field of the separable polynomial $f(x) \in F[x]$ then we may also view $f(x)$ as an element of $E[x]$ for any subfield E of K containing F. Then K is also the splitting field of $f(x)$ over E, so the extension K/E is Galois. By Corollary 10, E is the fixed field of $\text{Aut}(K/E) \leq G$, showing that *every* subfield of K containing F arises as the fixed field for some subgroup of G. Hence the correspondence above is surjective from right to left, hence a bijection. The correspondences are inverse to each other since the automorphisms fixing E are precisely $\text{Aut}(K/E)$ by Corollary 10.

We have already seen that the Galois correspondence is inclusion reversing in Proposition 4, which gives (1).

If $E = K_H$ is the fixed field of H, then Theorem 9 gives $[K : E] = |H|$ and $[K : F] = |G|$. Taking the quotient gives $[E : F] = |G : H|$, which proves (2).

Corollary 11 gives (3) immediately.

Suppose $E = K_H$ is the fixed field of the subgroup H. Every $\sigma \in G = \text{Gal}(K/F)$ when restricted to E is an embedding $\sigma|_E$ of E with the subfield $\sigma(E)$ of K. Conversely, let $\tau : E \xrightarrow{\sim} \tau(E) \subseteq \overline{F}$ be any embedding of E (into a fixed algebraic closure $\overline{F}$ of F containing K) which fixes F. Then $\tau(E)$ is in fact contained in K: if $\alpha \in E$ has minimal polynomial $m_\alpha(x)$ over F then $\tau(\alpha)$ is another root of $m_\alpha(x)$ and K contains all these roots by Theorem 13. As above K is the splitting field of $f(x)$ over E and so also the splitting field of $\tau f(x)$ (which is the same as $f(x)$ since $f(x)$ has coefficients in F) over $\tau(E)$. Theorem 13.27 on extending isomorphisms then shows that we can extend τ to an isomorphism σ:

$$\begin{tikzcd} \sigma: & K \arrow[r, "\sim"] \arrow[d, dash] & K \arrow[d, dash] \\ \tau: & E \arrow[r, "\sim"] & \tau(E). \end{tikzcd}$$

Since σ fixes F (because τ does), it follows that *every* embedding τ of E fixing F is the restriction to E of some automorphism σ of K fixing F, in other words, every embedding of E is of the form $\sigma|_E$ for some $\sigma \in G$.

Two automorphisms $\sigma, \sigma' \in G$ restrict to the *same* embedding of E if and only if $\sigma^{-1}\sigma'$ is the identity map on E. But then $\sigma^{-1}\sigma' \in H$ (i.e., $\sigma' \in \sigma H$) since by (3) the automorphisms of K which fix E are precisely the elements in H. Hence the distinct embeddings of E are in bijection with the cosets σH of H in G. In particular this gives

$$|\text{Emb}(E/F)| = [G : H] = [E : F]$$

where $\text{Emb}(E/F)$ denotes the set of embeddings of E (into a fixed algebraic closure of F) which fix F. Note that $\text{Emb}(E/F)$ contains the automorphisms $\text{Aut}(E/F)$.

The extension E/F will be Galois if and only if $|\text{Aut}(E/F)| = [E : F]$. By the equality above, this will be the case if and only if each ofthe *embeddings* of E is actually an *automorphism* of E, i.e., if and only if $\sigma(E) = E$ for every $\sigma \in G$.

If $\sigma \in G$, then the subgroup of G fixing the field $\sigma(E)$ is the group $\sigma H \sigma^{-1}$, i.e.,

$$\sigma(E) = K_{\sigma H \sigma^{-1}}.$$

To see this observe that if $\sigma\alpha \in \sigma(E)$ then

$$(\sigma h \sigma^{-1})(\sigma\alpha) = \sigma(h\alpha) = \sigma\alpha \qquad \text{for all } h \in H,$$

since h fixes $\alpha \in E$, which shows that $\sigma H \sigma^{-1}$ fixes $\sigma(E)$. The group fixing $\sigma(E)$ has order equal to the degree of K over $\sigma(E)$. But this is the same as the degree of K over E since the fields are isomorphic, hence the same as the order of H. Hence $\sigma H \sigma^{-1}$ is precisely the group fixing $\sigma(E)$ since we have shown containment and their orders are the same.

Because of the bijective nature of the Galois correspondence already proved we know that two subfields of K containing F are equal if and only if their fixing subgroups are equal in G. Hence $\sigma(E) = E$ for all $\sigma \in G$ if and only if $\sigma H \sigma^{-1} = H$ for all $\sigma \in G$, in other words E is Galois over F if and only if H is a normal subgroup of G.

We have already identified the embeddings of E over F as the set of cosets of H in G and when H is normal in G seen that the embeddings are automorphisms. It follows that in this case the *group* of cosets G/H is identified with the *group* of automorphisms of the Galois extension E/F by the definition of the group operation (composition of automorphisms). Hence $G/H \cong \mathrm{Gal}(E/F)$ when H is normal in G, which completes the proof of (4).

Suppose H_1 is the subgroup of elements of G fixing the subfield E_1 and H_2 is the subgroup of elements of G fixing the subfield E_2. Any element in $H_1 \cap H_2$ fixes both E_1 and E_2, hence fixes every element in the composite E_1E_2, since the elements in this field are algebraic combinations of the elements of E_1 and E_2. Conversely, if an automorphism σ fixes the composite E_1E_2 then in particular σ fixes E_1, i.e., $\sigma \in H_1$, and σ fixes E_2, i.e., $\sigma \in H_2$, hence $\sigma \in H_1 \cap H_2$. This proves that the composite E_1E_2 corresponds to the intersection $H_1 \cap H_2$. Similarly, the intersection $E_1 \cap E_2$ corresponds to the group $\langle H_1, H_2 \rangle$ generated by H_1 and H_2, completing the proof of the theorem.

Example: ($\mathbb{Q}(\sqrt{2}, \sqrt{3})$ and $\mathbb{Q}(\sqrt[3]{2}, \rho)$)

We have already seen examples of this theorem at the beginning of this section. We now see that the diagrams of subfields for the two fields $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ and $\mathbb{Q}(\sqrt[3]{2}, \rho)$ given before indicate *all* the subfields for these two fields.

Since every subgroup of the Klein 4-group is normal, all the subfields of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ are Galois extensions of $\mathbb{Q}$.

Similarly, since the only nontrivial normal subgroup of S_3 is the subgroup of order 3, we see that only the subfield $\mathbb{Q}(\rho)$ of $K = \mathbb{Q}(\sqrt[3]{2}, \rho)$ is Galois over $\mathbb{Q}$, with Galois group isomorphic to $S_3/\langle \sigma \rangle$, i.e., the cyclic group of order 2. For example, the nontrivial automorphism of $\mathbb{Q}(\rho)$ is induced by restricting any element (τ, for instance) in the nontrivial coset of $\langle \sigma \rangle$ to $\mathbb{Q}(\rho)$. This is clear from the explicit descriptions of these automorphisms given before — each of the elements τ, $\tau\sigma$, $\tau\sigma^2$ in this coset map ρ to ρ^2. The restrictions of the elements of $\mathrm{Gal}(K/\mathbb{Q})$ to the (non-Galois) cubic subfields do not give automorphisms of these fields in general, rather giving isomorphisms of these fields with each other, in accordance with (4) of the theorem.

Example: ($\mathbb{Q}(\sqrt{2} + \sqrt{3})$)

Consider the field $\mathbb{Q}(\sqrt{2} + \sqrt{3})$. This is clearly a subfield of the Galois extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})$. The other roots of the minimal polynomial for $\sqrt{2} + \sqrt{3}$ over $\mathbb{Q}$ are therefore the distinct conjugates of $\sqrt{2}+\sqrt{3}$ under the Galois group. The conjugates are

$$\pm\sqrt{2}\pm\sqrt{3}$$

which are easily seen to be distinct. The minimal polynomial is therefore

$$[x-(\sqrt{2}+\sqrt{3})][x-(\sqrt{2}-\sqrt{3})][x-(-\sqrt{2}+\sqrt{3})][x-(-\sqrt{2}-\sqrt{3})]$$

which is quickly computed to be the polynomial x^4-10x^2+1. It follows that this polynomial is irreducible and that

$$\mathbb{Q}(\sqrt{2},\sqrt{3})=\mathbb{Q}(\sqrt{2}+\sqrt{3}),$$

either by degree considerations or by noting that only the automorphism 1 of $\{1,\sigma,\tau,\sigma\tau\}$ fixes $\sqrt{2}+\sqrt{3}$ so the fixing group for this field is the same as for $\mathbb{Q}(\sqrt{2},\sqrt{3})$.

Example: (Splitting Field of x^8-2)

The splitting field of x^8-2 over $\mathbb{Q}$ is generated by $\theta=\sqrt[8]{2}$ (any fixed 8th root of 2, say the real one) and a primitive 8th root of unity $\zeta=\zeta_8$. Recall from Section 13.6 that

$$\mathbb{Q}(\zeta_8)=\mathbb{Q}(i,\sqrt{2}).$$

Since $\theta^4=\sqrt{2}$ we see that the splitting field is generated by θ and i. The subfield $\mathbb{Q}(\theta)$ is of degree 8 over $\mathbb{Q}$ (since x^8-2 is irreducible, being Eisenstein), and all the elements of this field are real. Hence $i\notin\mathbb{Q}(\theta)$ and since i generates at most a quadratic extension of this field, the splitting field

$$\mathbb{Q}(\sqrt[8]{2},\zeta_8)=\mathbb{Q}(\sqrt[8]{2},i)$$

is of degree 16 over $\mathbb{Q}$.

The Galois group is determined by the action on the generators θ and i which gives the possibilities

$$\begin{cases}\theta\mapsto\zeta^a\theta & a=0,1,2,\dots,7\\ i\mapsto\pm i\end{cases}$$

Since we have already seen that the degree of the extension is 16 and there are only 16 possible such maps, it follows that in fact each of the maps above is an automorphism of $\mathbb{Q}(\sqrt[8]{2},i)$ over $\mathbb{Q}$.

Define the two automorphisms

$$\sigma:\begin{cases}\theta\mapsto\zeta\theta\\ i\mapsto i\end{cases}\qquad\qquad \tau:\begin{cases}\theta\mapsto\theta\\ i\mapsto -i\end{cases}$$

(τ is the map induced by complex conjugation). Since

$$\begin{aligned}\zeta=\zeta_8&=\frac{\sqrt{2}}{2}+i\frac{\sqrt{2}}{2}=\frac{1}{2}(1+i)\sqrt{2}\\ &=\frac{1}{2}(1+i)\theta^4\end{aligned}$$

we can easily compute what happens to ζ from the explicit expressions for the powers of ζ in the following Figure 1.

Using these explicit values we find

$$\sigma:\begin{cases}\theta\mapsto\zeta\theta\\ i\mapsto i\\ \zeta\mapsto-\zeta=\zeta^5\end{cases}\qquad\qquad \tau:\begin{cases}\theta\mapsto\theta\\ i\mapsto -i\\ \zeta\mapsto\zeta^7\end{cases}$$

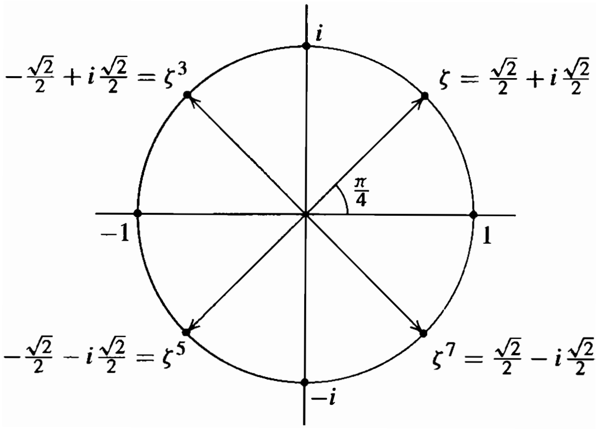

i
$-\frac{\sqrt{2}}{2}+i\frac{\sqrt{2}}{2}=\zeta^3$
$\zeta=\frac{\sqrt{2}}{2}+i\frac{\sqrt{2}}{2}$
$\frac{\pi}{4}$
-1
1
$-\frac{\sqrt{2}}{2}-i\frac{\sqrt{2}}{2}=\zeta^5$
$\zeta^7=\frac{\sqrt{2}}{2}-i\frac{\sqrt{2}}{2}$
$-i$

Fig. 1

Note that the reason we are interested in also keeping track of the action on the element ζ is that it will be needed in computing the composites of automorphisms, for example in computing

$$\begin{aligned}\sigma^2(\theta) &= \sigma(\zeta\theta) = \sigma(\zeta)\sigma(\theta) = (-\zeta)(\zeta\theta) = -\zeta^2\theta \\ &= -i\theta.\end{aligned}$$

We can similarly compute the following automorphisms:

$$\sigma : \begin{cases} \theta \mapsto \zeta\theta \\ i \mapsto i \\ \zeta \mapsto \zeta^5 \end{cases} \qquad \tau\sigma : \begin{cases} \theta \mapsto \zeta^7\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3 \end{cases}$$

$$\sigma^2 : \begin{cases} \theta \mapsto \zeta^6\theta \\ i \mapsto i \\ \zeta \mapsto \zeta \end{cases} \qquad \tau\sigma^2 : \begin{cases} \theta \mapsto \zeta^2\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^7 \end{cases}$$

$$\sigma^3 : \begin{cases} \theta \mapsto \zeta^7\theta \\ i \mapsto i \\ \zeta \mapsto -\zeta \end{cases} \qquad \tau\sigma^3 : \begin{cases} \theta \mapsto \zeta\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3 \end{cases}$$

$$\sigma^4 : \begin{cases} \theta \mapsto -\theta \\ i \mapsto i \\ \zeta \mapsto \zeta \end{cases} \qquad \tau\sigma^4 : \begin{cases} \theta \mapsto -\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^7 \end{cases}$$

$$\sigma^5 : \begin{cases} \theta \mapsto \zeta^5\theta \\ i \mapsto i \\ \zeta \mapsto -\zeta \end{cases} \qquad \tau\sigma^5 : \begin{cases} \theta \mapsto \zeta^3\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3 \end{cases}$$

$$\sigma^6 : \begin{cases} \theta \mapsto \zeta^2\theta \\ i \mapsto i \\ \zeta \mapsto \zeta \end{cases} \qquad \tau\sigma^6 : \begin{cases} \theta \mapsto \zeta^6\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^7 \end{cases}$$

$$\sigma^7 : \begin{cases} \theta \mapsto \zeta^3\theta \\ i \mapsto i \\ \zeta \mapsto -\zeta \end{cases} \qquad \tau\sigma^7 : \begin{cases} \theta \mapsto \zeta^5\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3. \end{cases}$$

Since this exhausts the possibilities, these elements (together with 1 and τ) are the Galois group. We see in particular that σ and τ generate the Galois group. To determine the relations satisfied by these elements, we observe first that clearly $\tau^2 = 1$ and $(\sigma^4)^2 = 1$, so that

$$\sigma^8 = \tau^2 = 1.$$

Also, we compute

$$\sigma\tau : \begin{cases} \theta \mapsto \zeta\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3 \end{cases}$$

so that

$$\sigma\tau = \tau\sigma^3.$$

It is not too difficult to show that these relations define the group completely, i.e.,

$$\mathrm{Gal}(\mathbb{Q}(\sqrt[8]{2}, i)/\mathbb{Q}) = \langle\ \sigma, \tau \mid \sigma^8 = \tau^2 = 1, \sigma\tau = \tau\sigma^3\ \rangle.$$

Such a group is called a *quasidihedral group* (recall that the dihedral group of order 16 would have the relation $\sigma\tau = \tau\sigma^7$ instead of $\sigma\tau = \tau\sigma^3$) and is a subgroup of S_8 since the Galois group is a subgroup of the permutations of the 8 roots of $x^8 - 2$.

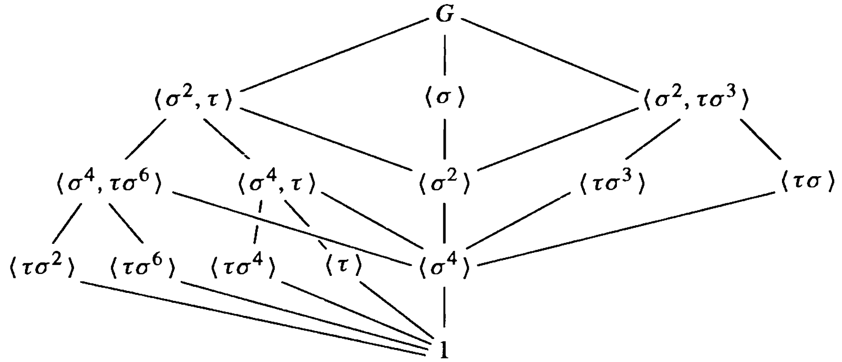
This example again illustrates that one must take care in determining Galois groups from the actions on generators. We first computed the degree of the Galois extension above to determine the number of elements in the Galois group. Had we proceeded directly from the original generators $\theta = \sqrt[8]{2}$ and $\zeta = \zeta_8$ we might have (incorrectly) concluded that there were a total of 32 elements in the Galois group, since the first generator is mapped to any of 8 possible roots of $x^8 - 2$ and the second generator is mapped to any of 4 possible roots of its minimal polynomial $\Phi_4(x) = x^4 + 1$. The problem, as previously indicated, is that these choices are not independent. Here the reason is provided by the algebraic relation

$$\theta^4 = \sqrt{2} = \zeta + \zeta^7$$

which shows that one cannot specify the images of θ and ζ independently — their images must again satisfy this algebraic relation. This relation is perhaps sufficiently subtle to serve as a caution against rashly concluding maps are automorphisms. We note that in general it is necessary to provide justification that maps are automorphisms. This can be accomplished for example by using the extension theorems or by using degree considerations as we did here.

Determining the lattice of subgroups of this group G is a straightforward problem.

The lattice is the following:


G
$\langle \sigma^2, \tau \rangle$
$\langle \sigma \rangle$
$\langle \sigma^2, \tau\sigma^3 \rangle$
$\langle \sigma^4, \tau\sigma^6 \rangle$
$\langle \sigma^4, \tau \rangle$
$\langle \sigma^2 \rangle$
$\langle \tau\sigma^3 \rangle$
$\langle \tau\sigma \rangle$
$\langle \tau\sigma^2 \rangle$
$\langle \tau\sigma^6 \rangle$
$\langle \tau\sigma^4 \rangle$
$\langle \tau \rangle$
$\langle \sigma^4 \rangle$
1

Determining the subfields corresponding to these subgroups (which by the Fundamental Theorem gives *all* the subfields of $\mathbb{Q}(\sqrt[8]{2}, i))$ is quite simple for a number of the subgroups above using (2) of the Fundamental Theorem, which states that the degree of the extension over $\mathbb{Q}$ is equal to the *index* of the fixing subgroup. It then suffices to find a subfield of the right degree which is fixed by the subgroup in question. Remember also that if a subfield is fixed by the *generators* of a subgroup, then it is fixed by the subgroup. For example, from the explicit description for the automorphism σ we see that $\mathbb{Q}(i)$ is fixed by the group generated by σ. Since this is a subgroup of index 2 and $\mathbb{Q}(i)$ is of degree 2 over $\mathbb{Q}$, it must be the full fixed field. Most of the fixed fields for the subgroups above can be determined in as simple a manner.

For the subgroups of order 4 on the right (namely, generated by $\tau\sigma^3$ and by $\tau\sigma$), it is perhaps not so easy to see how to determine the corresponding fixed field. For the subgroup H generated by $\tau\sigma^3$ we may proceed as follows: the element $\theta^2 = \sqrt[4]{2}$ is clearly fixed by σ^4. By the diagram above, σ^4 is a normal subgroup of H of index 2, with representatives $1, \tau\sigma^3$ for the cosets. Consider the element

$$\alpha = (1 + \tau\sigma^3)\theta^2 = \theta^2 + \tau\sigma^3\theta^2.$$

Then α is fixed by σ^4 (we are in a commutative group H of order 4, so σ^4 commutes with 1 and $\tau\sigma^3$ and we already know θ^2 is fixed by σ^4). But (and this is the point), α is also fixed by $\tau\sigma^3$:

$$\begin{aligned}\tau\sigma^3\alpha &= \tau\sigma^3(1 + \tau\sigma^3)\theta^2 = [\tau\sigma^3 + (\tau\sigma^3)^2]\theta^2 \\ &= (\tau\sigma^3 + \sigma^4)\theta^2\end{aligned}$$

and the last expression is just α since $\sigma^4\theta^2 = \theta^2$. Hence α is an element of the fixed field for H. Explicitly

$$\alpha = \sqrt[4]{2} + i\sqrt[4]{2} = (1 + i)\sqrt[4]{2}.$$

A quick check shows that α is not fixed by the automorphism σ^2, so by the diagram of subgroups above, it follows that the fixing subgroup for the field $\mathbb{Q}(\alpha)$ is no larger than H, hence is precisely H, which gives us our fixed field. This also gives the fixed field for $\langle \tau\sigma \rangle$ by recalling that in general if E is the fixed field of H then the fixed field of $\tau H\tau^{-1}$ is the field $\tau(E)$. For $H = \langle \tau\sigma^3 \rangle, \tau H\tau^{-1} = \langle \tau\sigma \rangle$, with fixed field given by $\tau(\alpha) = (1 - i)\sqrt[4]{2}$.

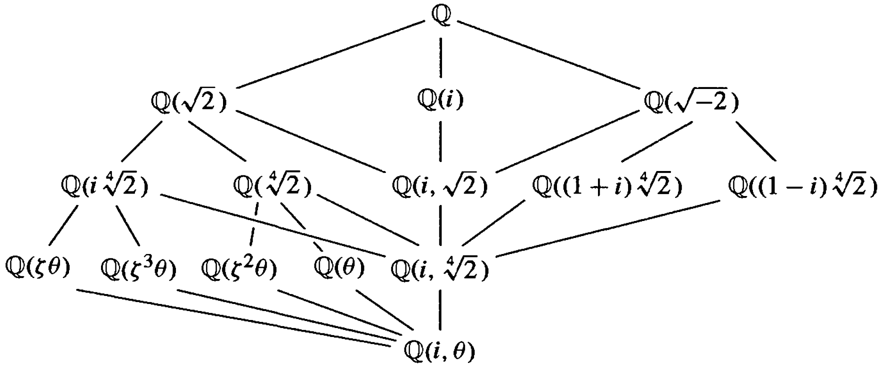
In general one tries to determine elements which are fixed by a given subgroup H of the Galois group (cf. the exercises, which indicate where the element above arose) and attempts to generate a sufficiently large field to give the full fixed field. In our case we were able to accomplish this with a single generator. We shall see later that every finite extension of $\mathbb{Q}$ is a simple extension, so there will be a single generator of this type, but in general it may be difficult to produce it directly.

The element α is a root of the polynomial

$$x^4 + 8$$

which must therefore be irreducible since we have already determined that a root of this polynomial generates an extension of degree 4 over $\mathbb{Q}$.

In a similar way it is possible to complete the diagram of subfields of $\mathbb{Q}(\sqrt[8]{2}, i)$, which we have inverted to emphasize its relation with the subgroup diagram above ($\theta = \sqrt[8]{2}$):



$\mathbb{Q}$
$\mathbb{Q}(\sqrt{2})$
$\mathbb{Q}(i)$
$\mathbb{Q}(\sqrt{-2})$
$\mathbb{Q}(i\sqrt[4]{2})$
$\mathbb{Q}(\sqrt[4]{2})$
$\mathbb{Q}(i, \sqrt{2})$
$\mathbb{Q}((1+i)\sqrt[4]{2})$
$\mathbb{Q}((1-i)\sqrt[4]{2})$
$\mathbb{Q}(\zeta\theta)$
$\mathbb{Q}(\zeta^3\theta)$
$\mathbb{Q}(\zeta^2\theta)$
$\mathbb{Q}(\theta)$
$\mathbb{Q}(i, \sqrt[4]{2})$
$\mathbb{Q}(i, \theta)$

Note that the group $\langle \sigma^4 \rangle$ is normal in G (in fact it is the center of G) with quotient $G/\langle \sigma^4 \rangle \cong D_8$, so the corresponding fixed field $\mathbb{Q}(i, \sqrt[4]{2})$ is Galois over $\mathbb{Q}$ with D_8 as Galois group. Being Galois it is a splitting field, evidently the splitting field for $x^4 - 2$. The lattice of subfields for this field is then immediate from the lattice above.

We end this example with the following amusing aspect of this Galois extension. It is an easy exercise to verify that

$$\langle \sigma^2, \tau \rangle \cong D_8 \quad \langle \sigma \rangle \cong \mathbb{Z}/8\mathbb{Z} \quad \langle \sigma^2, \tau\sigma^3 \rangle \cong Q_8$$

where D_8 is the dihedral group of order 8 and Q_8 is the quaternion group of order 8. It follows that the field $\mathbb{Q}(\sqrt[8]{2}, i)$ is Galois of degree 8 over its three quadratic subfields

$$\mathbb{Q}(\sqrt{2}) \qquad \mathbb{Q}(i) \qquad \mathbb{Q}(\sqrt{-2})$$

with dihedral, cyclic and quaternion Galois groups, respectively, so that three of the 5 possible groups of order 8 (and both non-abelian ones) appear as Galois groups in this extension.

We shall consider additional examples and applications in the following sections.

EXERCISES

1. Determine the minimal polynomial over $\mathbb{Q}$ for the element $\sqrt{2} + \sqrt{5}$.
2. Determine the minimal polynomial over $\mathbb{Q}$ for the element $1 + \sqrt[3]{2} + \sqrt[3]{4}$.
3. Determine the Galois group of $(x^2 - 2)(x^2 - 3)(x^2 - 5)$. Determine *all* the subfields of the splitting field of this polynomial.

4. Let p be a prime. Determine the elements of the Galois group of $x^p - 2$.

5. Prove that the Galois group of $x^p - 2$ for p a prime is isomorphic to the group of matrices $\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$ where $a, b \in \mathbb{F}_p$, $a \neq 0$.

6. Let $K = \mathbb{Q}(\sqrt[8]{2}, i)$ and let $F_1 = \mathbb{Q}(i)$, $F_2 = \mathbb{Q}(\sqrt{2})$, $F_3 = \mathbb{Q}(\sqrt{-2})$. Prove that $\text{Gal}(K/F_1) \cong Z_8$, $\text{Gal}(K/F_2) \cong D_8$, $\text{Gal}(K/F_3) \cong Q_8$.

7. Determine all the subfields of the splitting field of $x^8 - 2$ which are Galois over $\mathbb{Q}$.

8. Suppose K is a Galois extension of F of degree p^n for some prime p and some $n \geq 1$. Show there are Galois extensions of F contained in K of degrees p and p^{n-1}.

9. Give an example of fields F_1, F_2, F_3 with $\mathbb{Q} \subset F_1 \subset F_2 \subset F_3$, $[F_3 : \mathbb{Q}] = 8$ and each field is Galois over all its subfields with the exception that F_2 is not Galois over $\mathbb{Q}$.

10. Determine the Galois group of the splitting field over $\mathbb{Q}$ of $x^8 - 3$.

11. Suppose $f(x) \in \mathbb{Z}[x]$ is an irreducible quartic whose splitting field has Galois group S_4 over $\mathbb{Q}$ (there are many such quartics, cf. Section 6). Let θ be a root of $f(x)$ and set $K = \mathbb{Q}(\theta)$. Prove that K is an extension of $\mathbb{Q}$ of degree 4 which has no proper subfields. Are there any Galois extensions of $\mathbb{Q}$ of degree 4 with no proper subfields?

12. Determine the Galois group of the splitting field over $\mathbb{Q}$ of $x^4 - 14x^2 + 9$.

13. Prove that if the Galois group of the splitting field of a cubic over $\mathbb{Q}$ is the cyclic group of order 3 then all the roots of the cubic are real.

14. Show that $\mathbb{Q}(\sqrt{2 + \sqrt{2}})$ is a cyclic quartic field, i.e., is a Galois extension of degree 4 with cyclic Galois group.

15. (*Biquadratic Extensions*) Let F be a field of characteristic $\neq 2$.
   (a) If $K = F(\sqrt{D_1}, \sqrt{D_2})$ where $D_1, D_2 \in F$ have the property that none of D_1, D_2 or $D_1 D_2$ is a square in F, prove that K/F is a Galois extension with $\text{Gal}(K/F)$ isomorphic to the Klein 4-group.
   (b) Conversely, suppose K/F is a Galois extension with $\text{Gal}(K/F)$ isomorphic to the Klein 4-group. Prove that $K = F(\sqrt{D_1}, \sqrt{D_2})$ where $D_1, D_2 \in F$ have the property that none of D_1, D_2 or $D_1 D_2$ is a square in F.

16. (a) Prove that $x^4 - 2x^2 - 2$ is irreducible over $\mathbb{Q}$.
   (b) Show the roots of this quartic are

$$\alpha_1 = \sqrt{1 + \sqrt{3}} \qquad \alpha_3 = -\sqrt{1 + \sqrt{3}}$$
$$\alpha_2 = \sqrt{1 - \sqrt{3}} \qquad \alpha_4 = -\sqrt{1 - \sqrt{3}}.$$

   (c) Let $K_1 = \mathbb{Q}(\alpha_1)$ and $K_2 = \mathbb{Q}(\alpha_2)$. Show that $K_1 \neq K_2$, and $K_1 \cap K_2 = \mathbb{Q}(\sqrt{3}) = F$.
   (d) Prove that K_1, K_2 and $K_1 K_2$ are Galois over F with $\text{Gal}(K_1 K_2/F)$ the Klein 4-group. Write out the elements of $\text{Gal}(K_1 K_2/F)$ explicitly. Determine all the subgroups of the Galois group and give their corresponding fixed subfields of $K_1 K_2$ containing F.
   (e) Prove that the splitting field of $x^4 - 2x^2 - 2$ over $\mathbb{Q}$ is of degree 8 with dihedral Galois group.

The following two exercises indicate one method for constructing elements in subfields of a given field and are quite useful in many computations.

17. Let K/F be any finite extension and let $\alpha \in K$. Let L be a Galois extension of F containing K and let $H \leq \text{Gal}(L/F)$ be the subgroup corresponding to K. Define the *norm* of α from

K to F to be

$$\mathrm{N}_{K/F}(\alpha) = \prod_{\sigma} \sigma(\alpha),$$

where the product is taken over all the embeddings of K into an algebraic closure of F (so over a set of coset representatives for H in $\mathrm{Gal}(L/F)$ by the Fundamental Theorem of Galois Theory). This is a product of Galois conjugates of α. In particular, if K/F is Galois this is $\prod_{\sigma \in \mathrm{Gal}(K/F)} \sigma(\alpha)$.

(a) Prove that $\mathrm{N}_{K/F}(\alpha) \in F$.

(b) Prove that $\mathrm{N}_{K/F}(\alpha\beta) = \mathrm{N}_{K/F}(\alpha)\mathrm{N}_{K/F}(\beta)$, so that the norm is a multiplicative map from K to F.

(c) Let $K = F(\sqrt{D})$ be a quadratic extension of F. Show that $\mathrm{N}_{K/F}(a + b\sqrt{D}) = a^2 - Db^2$.

(d) Let $m_\alpha(x) = x^d + a_{d-1}x^{d-1} + \cdots + a_1 x + a_0 \in F[x]$ be the minimal polynomial for $\alpha \in K$ over F. Let $n = [K : F]$. Prove that d divides n, that there are d distinct Galois conjugates of α which are all repeated n/d times in the product above and conclude that $\mathrm{N}_{K/F}(\alpha) = (-1)^n a_0^{n/d}$.

18. With notation as in the previous problem, define the *trace* of α from K to F to be

$$\mathrm{Tr}_{K/F}(\alpha) = \sum_{\sigma} \sigma(\alpha),$$

a sum of Galois conjugates of α.

(a) Prove that $\mathrm{Tr}_{K/F}(\alpha) \in F$.

(b) Prove that $\mathrm{Tr}_{K/F}(\alpha + \beta) = \mathrm{Tr}_{K/F}(\alpha) + \mathrm{Tr}_{K/F}(\beta)$, so that the trace is an additive map from K to F.

(c) Let $K = F(\sqrt{D})$ be a quadratic extension of F. Show that $\mathrm{Tr}_{K/F}(a + b\sqrt{D}) = 2a$.

(d) Let $m_\alpha(x)$ be as in the previous problem. Prove that $\mathrm{Tr}_{K/F}(\alpha) = -\dfrac{n}{d} a_{d-1}$.

19. With notation as in the previous problems show that $\mathrm{N}_{K/F}(a\alpha) = a^n \mathrm{N}_{K/F}(\alpha)$ and $\mathrm{Tr}_{K/F}(a\alpha) = a\mathrm{Tr}_{K/F}(\alpha)$ for all a in the base field F. In particular show that $\mathrm{N}_{K/F}(a) = a^n$ and $\mathrm{Tr}_{K/F}(a) = na$ for all $a \in F$.

20. With notation as in the previous problems show more generally that $\prod_\sigma (x - \sigma(\alpha)) = (m_\alpha(x))^{n/d}$.

21. Use the linear independence of characters to show that for any Galois extension K of F there is an element $\alpha \in K$ with $\mathrm{Tr}_{K/F}(\alpha) \neq 0$.

22. Suppose K/F is a Galois extension and let σ be an element of the Galois group.

(a) Suppose $\alpha \in K$ is of the form $\alpha = \dfrac{\beta}{\sigma\beta}$ for some nonzero $\beta \in K$. Prove that $\mathrm{N}_{K/F}(\alpha) = 1$.

(b) Suppose $\alpha \in K$ is of the form $\alpha = \beta - \sigma\beta$ for some $\beta \in K$. Prove that $\mathrm{Tr}_{K/F}(\alpha) = 0$.

The next exercise and Exercise 26 following establish the multiplicative and additive forms of Hilbert's Theorem 90. These are instances of the vanishing of a first cohomology group, as will be discussed in Section 17.3.

23. (*Hilbert's Theorem 90*) Let K be a Galois extension of F with cyclic Galois group of order n generated by σ. Suppose $\alpha \in K$ has $\mathrm{N}_{K/F}(\alpha) = 1$. Prove that α is of the form $\alpha = \dfrac{\beta}{\sigma\beta}$ for some nonzero $\beta \in K$. [By the linear independence of characters show there exists some $\theta \in K$ such that

$$\beta = \theta + \alpha\sigma(\theta) + (\alpha\,\sigma\alpha)\sigma^2(\theta) + \cdots + (\alpha\,\sigma\alpha \ldots \sigma^{n-2}\alpha)\sigma^{n-1}(\theta)$$

is nonzero. Compute $\frac{\beta}{\sigma\beta}$ using the fact that α has norm 1 to F.]

24. Prove that the rational solutions $a, b \in \mathbb{Q}$ of Pythagoras' equation $a^2 + b^2 = 1$ are of the form $a = \frac{s^2 - t^2}{s^2 + t^2}$ and $b = \frac{2st}{s^2 + t^2}$ for some $s, t \in \mathbb{Q}$ and hence show that any right triangle with integer sides has sides of lengths $(m^2 - n^2, 2mn, m^2 + n^2)$ for some integers m, n. [Note that $a^2 + b^2 = 1$ is equivalent to $\mathrm{N}_{\mathbb{Q}(i)/\mathbb{Q}}(a + ib) = 1$, then use Hilbert's Theorem 90 above with $\beta = s + it$.]

25. Generalize the previous problem to determine all the rational solutions of the equation $a^2 + Db^2 = 1$ for $D \in \mathbb{Z}$, $D > 0$, D not a perfect square in $\mathbb{Z}$.

26. (*Additive Hilbert's Theorem 90*) Let K be a Galois extension of F with cyclic Galois group of order n generated by σ. Suppose $\alpha \in K$ has $\mathrm{Tr}_{K/F}(\alpha) = 0$. Prove that α is of the form $\alpha = \beta - \sigma\beta$ for some $\beta \in K$. [Let $\theta \in K$ be an element with $\mathrm{Tr}_{K/F}(\theta) \neq 0$ by a previous exercise, let

$$\beta = \frac{1}{\mathrm{Tr}_{K/F}(\theta)}[\alpha\sigma(\theta) + (\alpha + \sigma\alpha)\sigma^2(\theta) + \cdots + (\alpha + \sigma\alpha + \cdots + \sigma^{n-2}\alpha)\sigma^{n-1}(\theta)]$$

and compute $\beta - \sigma\beta$.]

27. Let $\alpha = \sqrt{(2 + \sqrt{2})(3 + \sqrt{3})}$ (positive real square roots for concreteness) and consider the extension $E = \mathbb{Q}(\alpha)$.

(a) Show that $a = (2 + \sqrt{2})(3 + \sqrt{3})$ is not a square in $F = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. [If $a = c^2$, $c \in F$, then $a\varphi(a) = (2 + \sqrt{2})^2(6) = (c\varphi c)^2$ for the automorphism $\varphi \in \mathrm{Gal}(F/\mathbb{Q})$ fixing $\mathbb{Q}(\sqrt{2})$. Since $c\varphi c = \mathrm{N}_{F/\mathbb{Q}(\sqrt{2})}(c) \in \mathbb{Q}(\sqrt{2})$ conclude that this implies $\sqrt{6} \in \mathbb{Q}(\sqrt{2})$, a contradiction.]

(b) Conclude from (a) that $[E : \mathbb{Q}] = 8$. Prove that the roots of the minimal polynomial over $\mathbb{Q}$ for α are the 8 elements $\pm\sqrt{(2 \pm \sqrt{2})(3 \pm \sqrt{3})}$.

(c) Let $\beta = \sqrt{(2 - \sqrt{2})(3 + \sqrt{3})}$. Show that $\alpha\beta = \sqrt{2}(3 + \sqrt{3}) \in F$ so that $\beta \in E$. Show similarly that the other roots are also elements of E so that E is a Galois extension of $\mathbb{Q}$. Show that the elements of the Galois group are precisely the maps determined by mapping α to one of the eight elements in (b).

(d) Let $\sigma \in \mathrm{Gal}(E/\mathbb{Q})$ be the automorphism which maps α to β. Show that since $\sigma(\alpha^2) = \beta^2$ that $\sigma(\sqrt{2}) = -\sqrt{2}$ and $\sigma(\sqrt{3}) = \sqrt{3}$. From $\alpha\beta = \sqrt{2}(3 + \sqrt{3})$ conclude that $\sigma(\alpha\beta) = -\alpha\beta$ and hence $\sigma(\beta) = -\alpha$. Show that σ is an element of order 4 in $\mathrm{Gal}(E/\mathbb{Q})$.

(e) Show similarly that the map τ defined by $\tau(\alpha) = \sqrt{(2 + \sqrt{2})(3 - \sqrt{3})}$ is an element of order 4 in $\mathrm{Gal}(E/\mathbb{Q})$. Prove that σ and τ generate the Galois group, $\sigma^4 = \tau^4 = 1$, $\sigma^2 = \tau^2$ and that $\sigma\tau = \tau\sigma^3$.

(f) Conclude that $\mathrm{Gal}(E/\mathbb{Q}) \cong Q_8$, the quaternion group of order 8.

28. Let $f(x) \in F[x]$ be an irreducible polynomial of degree n over the field F, let L be the splitting field of $f(x)$ over F and let α be a root of $f(x)$ in L. If K is any Galois extension of F contained in L, show that the polynomial $f(x)$ splits into a product of m irreducible polynomials each of degree d over K, where $m = [F(\alpha) \cap K : F]$ and $d = [K(\alpha) : K]$ (cf. also the generalization in Exercise 4 of Section 4). [If H is the subgroup of the Galois group of L over F corresponding to K then the factors of $f(x)$ over K correspond to the orbits of H on the roots of $f(x)$. Then use Exercise 9 of Section 4.1.]

29. Let k be a field and let $k(t)$ be the field of rational functions in the variable t. Define the maps σ and τ of $k(t)$ to itself by $\sigma f(t) = f(\frac{1}{1-t})$ and $\tau f(t) = f(\frac{1}{t})$ for $f(t) \in k(t)$.

(a) Prove that σ and τ are automorphisms of $k(t)$ (cf. Exercise 8 of Section 1) and that the group $G = \langle \sigma, \tau \rangle$ they generate is isomorphic to S_3.

(b) Prove that the element $t = \dfrac{(t^2 - t + 1)^3}{t^2(t-1)^2}$ is fixed by all the elements of G.

(c) Prove that $k(t)$ is precisely the fixed field of G in $k(t)$ [compute the degree of the extension].

30. Prove that the fixed field of the subgroup of automorphisms generated by τ in the previous problem is $k(t + \frac{1}{t})$. Prove that the fixed field of the subgroup generated by the automorphism $\tau\sigma^2$ (which maps t to $1-t$) is $k(t(1-t))$. Determine the fixed field of the subgroup generated by $\tau\sigma$ and the fixed field of the subgroup generated by σ.

31. Let K be a finite extension of F of degree n. Let α be an element of K.

(a) Prove that α acting by left multiplication on K is an F-linear transformation T_α of K.

(b) Prove that the minimal polynomial for α over F is the same as the minimal polynomial for the linear transformation T_α.

(c) Prove that the trace $\text{Tr}_{K/F}(\alpha)$ is the trace of the $n \times n$ matrix defined by T_α (which justifies these two uses of the same word "trace"). Prove that the norm $\text{N}_{K/F}(\alpha)$ is the determinant of T_α.

14.3 FINITE FIELDS

A finite field $\mathbb{F}$ has characteristic p for some prime p so is a finite dimensional vector space over $\mathbb{F}_p$. If the dimension is n, i.e., $[\mathbb{F} : \mathbb{F}_p] = n$, then $\mathbb{F}$ has precisely p^n elements. We have already seen (following Proposition 13.37) that $\mathbb{F}$ is then isomorphic to the splitting field of the polynomial $x^{p^n} - x$, hence is unique up to isomorphism. We denote the finite field of order p^n by $\mathbb{F}_{p^n}$.

The field $\mathbb{F}_{p^n}$ is Galois over $\mathbb{F}_p$, with cyclic Galois group of order n generated by the Frobenius automorphism

$$\text{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p) = \langle \sigma_p \rangle \cong \mathbb{Z}/n\mathbb{Z}$$

where

$$\begin{aligned} \sigma_p : \mathbb{F}_{p^n} &\to \mathbb{F}_{p^n} \\ \alpha &\mapsto \alpha^p \end{aligned}$$

(Example 7 following Corollary 6). By the Fundamental Theorem, every subfield of $\mathbb{F}_{p^n}$ corresponds to a subgroup of $\mathbb{Z}/n\mathbb{Z}$. Hence for every divisor d of n there is precisely one subfield of $\mathbb{F}_{p^n}$ of degree d over $\mathbb{F}_p$, namely the fixed field of the subgroup generated by σ_p^d of order n/d, and there are no other subfields. This field is isomorphic to $\mathbb{F}_{p^d}$, the unique finite field of order p^d.

Since the Galois group is abelian, every subgroup is normal, so each of the subfields $\mathbb{F}_{p^d}$ (d a divisor of n) is Galois over $\mathbb{F}_p$ (which is also clear from the fact that these are themselves splitting fields). Further, the Galois group $\text{Gal}(\mathbb{F}_{p^d}/\mathbb{F}_p)$ is generated by the image of σ_p in the quotient group $\text{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)/\langle \sigma_p^d \rangle$. If we denote this element again by σ_p, we recover the Frobenius automorphism for the extension $\mathbb{F}_{p^d}/\mathbb{F}_p$. (Note, however, that σ_p has order n in $\text{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$ and order d in $\text{Gal}(\mathbb{F}_{p^d}/\mathbb{F}_p)$.)

We summarize this in the following proposition.

Proposition 15. Any finite field is isomorphic to $\mathbb{F}_{p^n}$ for some prime p and some integer $n \geq 1$. The field $\mathbb{F}_{p^n}$ is the splitting field over $\mathbb{F}_p$ of the polynomial $x^{p^n} - x$, with cyclic Galois group of order n generated by the Frobenius automorphism σ_p. The subfields of $\mathbb{F}_{p^n}$ are all Galois over $\mathbb{F}_p$ and are in one to one correspondence with the divisors d of n. They are the fields $\mathbb{F}_{p^d}$, the fixed fields of ${\sigma_p}^d$.

The corresponding statements for the finite extensions of any finite field are easy consequences of Proposition 15 and are outlined in the exercises.

As an elementary application we have the following result on the polynomial x^4+1 in $\mathbb{Z}[x]$.

Corollary 16. The irreducible polynomial $x^4 + 1 \in \mathbb{Z}[x]$ is reducible modulo every prime p.

Proof: Consider the polynomial $x^4 + 1$ over $\mathbb{F}_p[x]$ for the prime p. If $p = 2$ we have $x^4 + 1 = (x + 1)^4$ and the polynomial is reducible. Assume now that p is odd. Then $p^2 - 1$ is divisible by 8 since p is congruent mod 8 to 1, 3, 5 or 7 and all of these square to 1 mod 8. Hence $x^{p^2-1} - 1$ is divisible by $x^8 - 1$. Then we have the divisibilities

$$x^4 + 1 \mid x^8 - 1 \mid x^{p^2-1} - 1 \mid x^{p^2} - x$$

which shows that all the roots of $x^4 + 1$ are roots of $x^{p^2} - x$. (Equivalently, these roots are fixed by the square of the Frobenius automorphism σ_p^2.) Since the roots of $x^{p^2} - x$ are the field $\mathbb{F}_{p^2}$, it follows that the extension generated by any root of $x^4 + 1$ is at most of degree 2 over $\mathbb{F}_p$, which means that $x^4 + 1$ cannot be irreducible over $\mathbb{F}_p$.

The multiplicative group ${\mathbb{F}_{p^n}}^\times$ is obviously a finite subgroup of the multiplicative group of a field. By Proposition 9.18, this is a *cyclic* group. If θ is any generator, then clearly $\mathbb{F}_{p^n} = \mathbb{F}_p(\theta)$. This proves the following result.

Proposition 17. The finite field $\mathbb{F}_{p^n}$ is simple. In particular, there exists an irreducible polynomial of degree n over $\mathbb{F}_p$ for every $n \geq 1$.

We have described the finite fields $\mathbb{F}_{p^n}$ above as the splitting fields of the polynomials $x^{p^n} - x$. By the previous proposition, this field can also be described as a quotient of $\mathbb{F}_p[x]$, namely by the minimal polynomial for θ. Since θ is necessarily a root of $x^{p^n} - x$, we see that the minimal polynomial for θ is a divisor of $x^{p^n} - x$ of degree n.

Conversely, let $p(x)$ be any irreducible polynomial of degree d, say, dividing $x^{p^n} - x$. If α is a root of $p(x)$, then the extension $\mathbb{F}_p(\alpha)$ is a subfield of $\mathbb{F}_{p^n}$ of degree d. Hence d is a divisor of n and the extension is Galois by Proposition 15 (in fact, the extension $\mathbb{F}_{p^d}$) so in particular all the roots of $p(x)$ are contained in $\mathbb{F}_p(\alpha)$.

The elements of $\mathbb{F}_{p^n}$ are precisely the roots of $x^{p^n} - x$. If we group together the factors $x - \alpha$ of this polynomial according to the degree d of their minimal polynomials over $\mathbb{F}_p$, we obtain

Proposition 18. The polynomial $x^{p^n} - x$ is precisely the product of all the distinct irreducible polynomials in $\mathbb{F}_p[x]$ of degree d where d runs through all divisors of n.

This proposition can be used to produce irreducible polynomials over $\mathbb{F}_p$ recursively. For example, the irreducible quadratics over $\mathbb{F}_2$ are the divisors of

$$\frac{x^4 - x}{x(x-1)}$$

which gives the single polynomial $x^2 + x + 1$. Similarly, the irreducible cubics over this field are the divisors of

$$\frac{x^8 - x}{x(x-1)} = x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$$

which factors into the two cubics $x^3 + x + 1$ and $x^3 + x^2 + 1$. The irreducible quartics are given by dividing $x^{16} - x$ by $x(x-1)$ and the irreducible quadratic $x^2 + x + 1$ above and then factoring into irreducible quartics:

$$\frac{x^{16} - x}{x(x-1)(x^2+x+1)} = (x^4 + x^3 + x^2 + x + 1)(x^4 + x^3 + 1)(x^4 + x + 1).$$

This gives a method for determining the product of all the irreducible polynomials over $\mathbb{F}_p$ of a given degree. There exist efficient algorithms for factorization of polynomials mod p which will give the individual irreducible polynomials (cf. the exercises) in practice. The importance of having irreducible polynomials at hand is that they give a representation of the finite fields $\mathbb{F}_{p^n}$ (as quotients $\mathbb{F}_p[x]/(f(x))$ for $f(x)$ irreducible of degree n) conducive to explicit computations.

Note also that since the finite field $\mathbb{F}_{p^n}$ is unique up to isomorphism, the quotients of $\mathbb{F}_p[x]$ by any of the irreducible polynomials of degree n are all isomorphic. If $f_1(x)$ and $f_2(x)$ are irreducible of degree n, then $f_2(x)$ splits completely in the field $\mathbb{F}_{p^n} \cong \mathbb{F}_p[x]/(f_1(x))$. If we denote a root of $f_2(x)$ by $\alpha(x)$ (to emphasize that it is a polynomial of degree $< n$ in x in $\mathbb{F}_p[x]/(f_1(x))$), then the isomorphism is given by

$$\begin{aligned} \mathbb{F}_p[x]/(f_2(x)) &\cong \mathbb{F}_p[x]/(f_1(x)) \\ x &\mapsto \alpha(x) \end{aligned}$$

(we have mapped a root of $f_2(x)$ in the first field to a root of $f_2(x)$ in the second field). For example, if $f_1(x) = x^4 + x^3 + 1$, $f_2(x) = x^4 + x + 1$ are two of the irreducible quartics over $\mathbb{F}_2$ determined above, then a simple computation verifies that

$$\alpha(x) = x^3 + x^2$$

is a root of $f_2(x)$ in $\mathbb{F}_{16} = \mathbb{F}_2[x]/(x^4 + x^3 + 1)$. Then we have

$$\begin{aligned} \mathbb{F}_2[x]/(x^4 + x + 1) &\cong \mathbb{F}_2[x]/(x^4 + x^3 + 1) \quad (\cong \mathbb{F}_{16}) \\ x &\mapsto x^3 + x^2. \end{aligned}$$

If we assume a result from elementary number theory we can give a formula for the number of irreducible polynomials of degree n. Define the *Möbius* μ-function by

$$\mu(n) = \begin{cases} 1 & \text{for } n = 1 \\ 0 & \text{if } n \text{ has a square factor} \\ (-1)^r & \text{if } n \text{ has } r \text{ distinct prime factors.} \end{cases}$$

If now $f(n)$ is a function defined for all nonnegative integers n and $F(n)$ is defined by

$$F(n) = \sum_{d|n} f(d) \qquad n = 1, 2, \ldots$$

then the *Möbius inversion formula* states that one can recover the function $f(n)$ from $F(n)$:

$$f(n) = \sum_{d|n} \mu(d)F(\frac{n}{d}) \qquad n = 1, 2, \ldots.$$

This is an elementary result from number theory which we take for granted. Define

$$\psi(n) = \text{the number of irreducible polynomials of degree } n \text{ in } \mathbb{F}_p[x].$$

Counting degrees in Proposition 18 we have

$$p^n = \sum_{d|n} d\psi(d).$$

Applying the Möbius inversion formula (for $f(n) = n\psi(n)$) we obtain

$$n\psi(n) = \sum_{d|n} \mu(d)p^{n/d}$$

which gives us a formula for the number of irreducible polynomials of degree n over $\mathbb{F}_p$:

$$\psi(n) = \frac{1}{n}\sum_{d|n} \mu(d)p^{n/d}.$$

For example, in the case $p = 2$, $n = 4$ we have

$$\psi(4) = \frac{1}{4}[\mu(1)2^4 + \mu(2)2^2 + \mu(4)2^1] = \frac{1}{4}(16 - 4 + 0) = 3$$

as we determined directly above.

We have seen above that

$$\mathbb{F}_{p^m} \subseteq \mathbb{F}_{p^n} \text{ if and only if } m \text{ divides } n.$$

In particular, given any two finite fields $\mathbb{F}_{p^{n_1}}$ and $\mathbb{F}_{p^{n_2}}$ there is a third finite field containing (an isomorphic copy of) them, namely $\mathbb{F}_{p^{n_1 n_2}}$. This gives us a partial ordering on these fields and allows us to think of their union. Since these give *all* the finite extensions of $\mathbb{F}_p$, we see that the union of $\mathbb{F}_{p^n}$ for all n is an algebraic closure of $\mathbb{F}_p$, unique up to isomorphism:

$$\overline{\mathbb{F}_p} = \bigcup_{n\geq 1} \mathbb{F}_{p^n}.$$

This provides a simple description of the algebraic closure of $\mathbb{F}_p$.

EXERCISES

1. Factor $x^8 - x$ into irreducibles in $\mathbb{Z}[x]$ and in $\mathbb{F}_2[x]$.
2. Write out the multiplication table for $\mathbb{F}_4$ and $\mathbb{F}_8$.
3. Prove that an algebraically closed field must be infinite.
4. Construct the finite field of 16 elements and find a generator for the multiplicative group. How many generators are there?
5. Exhibit an explicit isomorphism between the splitting fields of $x^3 - x + 1$ and $x^3 - x - 1$ over $\mathbb{F}_3$.
6. Suppose $K = \mathbb{Q}(\theta) = \mathbb{Q}(\sqrt{D_1}\ , \sqrt{D_2}\ )$ with $D_1, D_2 \in \mathbb{Z}$, is a biquadratic extension and that $\theta = a + b\sqrt{D_1} + c\sqrt{D_2} + d\sqrt{D_1 D_2}$ where $a, b, c, d \in \mathbb{Z}$ are integers. Prove that the minimal polynomial $m_\theta(x)$ for θ over $\mathbb{Q}$ is irreducible of degree 4 over $\mathbb{Q}$ but is reducible modulo every prime p. In particular show that the polynomial $x^4 - 10x^2 + 1$ is irreducible in $\mathbb{Z}[x]$ but is reducible modulo every prime. [Use the fact that there are no biquadratic extensions over finite fields.]
7. Prove that one of 2, 3 or 6 is a square in $\mathbb{F}_p$ for every prime p. Conclude that the polynomial

$$x^6 - 11x^4 + 36x^2 - 36 = (x^2 - 2)(x^2 - 3)(x^2 - 6)$$

has a root modulo p for every prime p but has no root in $\mathbb{Z}$.

8. Determine the splitting field of the polynomial $x^p - x - a$ over $\mathbb{F}_p$ where $a \neq 0$, $a \in \mathbb{F}_p$. Show explicitly that the Galois group is cyclic. [Show $\alpha \mapsto \alpha + 1$ is an automorphism.] Such an extension is called an *Artin–Schreier extension* (cf. Exercise 9 of Section 7).
9. Let $q = p^m$ be a power of the prime p and let $\mathbb{F}_q = \mathbb{F}_{p^m}$ be the finite field with q elements. Let $\sigma_q = \sigma_p^m$ be the m^{th} power of the Frobenius automorphism σ_p, called the q-Frobenius automorphism.
   (a) Prove that σ_q fixes $\mathbb{F}_q$.
   (b) Prove that every finite extension of $\mathbb{F}_q$ of degree n is the splitting field of $x^{q^n} - x$ over $\mathbb{F}_q$, hence is unique.
   (c) Prove that every finite extension of $\mathbb{F}_q$ of degree n is cyclic with σ_q as generator.
   (d) Prove that the subfields of the unique extension of $\mathbb{F}_q$ of degree n are in bijective correspondence with the divisors d of n.
10. Prove that n divides $\varphi(p^n - 1)$. [Observe that $\varphi(p^n - 1)$ is the order of the group of automorphisms of a cyclic group of order $p^n - 1$.]
11. Prove that $x^{p^n} - x + 1$ is irreducible over $\mathbb{F}_p$ only when $n = 1$ or $n = p = 2$. [Note that if α is a root, then so is $\alpha + a$ for any $a \in \mathbb{F}_{p^n}$. Show that this implies $\mathbb{F}_p(\alpha)$ contains $\mathbb{F}_{p^n}$ and that $[\mathbb{F}_p(\alpha) : \mathbb{F}_{p^n}] = p$.]

(*Berlekamp's Factorization Algorithm*) The following exercises outline the Berlekamp factorization algorithm for factoring polynomials in $\mathbb{F}_p[x]$. The efficiency of this algorithm is based on the efficiency of computing greatest common divisors in $\mathbb{F}_p[x]$ by the Euclidean Algorithm and on the efficiency of row-reduction matrix algorithms for solving systems of linear equations.

Let $f(x) \in \mathbb{F}_p[x]$ be a monic polynomial of degree n and let $f(x) = p_1(x)p_2(x)\dots p_k(x)$ where $p_1(x), p_2(x), \dots, p_k(x)$ are powers of distinct monic irreducibles in $\mathbb{F}_p[x]$.

12. Show that in order to write $f(x)$ as a product of irreducible polynomials in $\mathbb{F}_p[x]$ it suffices to determine the factors $p_1(x), \dots, p_k(x)$. [If $p(x) = q(x)^N \in \mathbb{F}_p[x]$ with $q(x)$ monic

and irreducible, show that $q(x)$ can be determined from $p(x)$ by checking for p^{th} powers and by computing greatest common divisors with derivatives.]

13. Let $g(x) \in \mathbb{F}_p[x]$ be any polynomial of degree $< n$. Denote by $R(h(x))$ the remainder of $h(x)$ after division by $f(x)$. Prove the following are equivalent:
 (a) $R(g(x^p)) = g(x)$.
 (b) $f(x)$ divides $[g(x)-0][g(x)-1]\dots[g(x)-(p-1)]$. [Use the fact that $g(x^p) = g(x)^p$ together with the factorization of $x^p - x$ in $\mathbb{F}_p[x]$.]
 (c) $p_i(x)$ divides the product in (b) for $i = 1, 2, \dots, k$.
 (d) For each i, $i = 1, 2, \dots, k$ there is an $s_i \in \mathbb{F}_p$ such that $p_i(x)$ divides $g(x) - s_i$, i.e., $g(x) \equiv s_i \pmod{p_i(x)}$.

14. Prove that the polynomials $g(x)$ of degree $< n$ satisfying the equivalent conditions of the previous exercise form a vector space V over $\mathbb{F}_p$ of dimension k. [Use the Chinese Remainder Theorem applied to the p^k possible choices for the s_i in 13(d)].

15. Let $g(x) = b_0 + b_1x + \cdots + b_{n-1}x^{n-1} \in V$. For $j = 0, 1, \dots, n-1$ let

$$R(x^{pj}) = a_{0,j} + a_{1,j}x + \cdots + a_{n-1,j}x^{n-1}$$

and let A be the $n \times n$ matrix

$$A = \begin{pmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0.n-1} \\ a_{1,0} & a_{1,1} & \cdots & a_{1.n-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n-1,0} & a_{n-1,1} & \cdots & a_{n-1,n-1} \end{pmatrix}. \qquad (*)$$

Show that condition (a) of Exercise 13 for $g(x) \in V$ is equivalent to

$$(A - I)B = 0 \qquad (**)$$

where B is the column matrix with entries $b_0, b_1, \dots, b_{n-1}$. Conclude that the rank of the matrix $A - I$ is $n - k$. Note that this already suffices to determine if $f(x)$ is irreducible, without actually determining the factors.

16. Let $g_1(x), g_2(x), \dots, g_k(x)$ be a basis of solutions to $(**)$ (so a basis for V), where we may take $g_1(x) = 1$. Beginning with $w(x) = f(x)$, compute the greatest common divisor $(w(x), g_i(x) - s)$ for $i = 2, 3, \dots, k$ and $s \in \mathbb{F}_p$ for every factor of $f(x)$ already computed. Note by Exercise 13(d) that every factor $p_i(x)$ of $f(x)$ divides such a g.c.d. The process terminates when k relatively prime factors have been determined.

Prove that this procedure actually gives all the factors $p_1(x), p_2(x), \dots, p_k(x)$, i.e., one can separate the individual factors $p_1(x), p_2(x), \dots, p_k(x)$ by this procedure, as follows:

If this were not the case, then for two of the factors, say $p_1(x)$ and $p_2(x)$, for each $i = 1, 2, \dots, k$ there would exist $s_i \in \mathbb{F}_p$ such that $g_i(x) - s_i$ is divisible by both $p_1(x)$ and $p_2(x)$. By the Chinese Remainder Theorem, choose a $g(x) \in V$ satisfying $g(x) \equiv 0 \pmod{p_1(x)}$ and $g(x) \equiv 1 \pmod{p_2(x)}$. Write $g(x) = \sum_{i=1}^{k} c_i g_i(x)$ in terms of the basis for V and let $s = \sum_{i=1}^{k} c_i s_i(x) \in \mathbb{F}_p$. Show that $s \equiv 0 \pmod{p_1(x)}$ so that $s = 0$ and $s \equiv 1 \pmod{p_2(x)}$ so that $s = 1$, a contradiction.

17. This exercise follows Berlekamp's Factorization Algorithm outlined in the previous exercises to determine the factorization of $f(x) = x^5 + x^2 + 4x + 6$ in $\mathbb{F}_7[x]$.
 (a) Show that $x^7 \equiv x^2 + 3x^3 + 6x^4 \pmod{f(x)}$. Similarly compute x^{14}, x^{21}, and x^{28} modulo $f(x)$ (note that x^{14} can most easily be computed by squaring the result for

x^7 and then reducing, etc.) to show that in this case the matrix A in Exercise 15 is

$$\begin{pmatrix} 1 & 0 & 5 & 1 & 4 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 1 & 3 & 3 & 3 \\ 0 & 3 & 4 & 2 & 2 \\ 0 & 6 & 3 & 1 & 1 \end{pmatrix}.$$

(b) Show that the reduced row echelon form for $A - I$ is the matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 6 \\ 0 & 0 & 1 & 0 & 6 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Conclude that $k = 2$ (so $f(x)$ is the product of precisely two factors which are powers of irreducible polynomials) and that $g_1(x) = 1$ and $g_2(x) = x^4 + 5x^3 + x^2 + x$ give a basis for the solutions to (**) in Exercise 15.

(c) Following the procedure in Exercise 16, show that $(f(x), g_2(x) - 1) = x^2 + 3x + 5 = p_1(x)$, with $f(x)/p_1(x) = x^3 + 4x^2 + 4x + 4 = p_2(x)$, giving the powers of the irreducible polynomials dividing $f(x)$ in $\mathbb{F}_7[x]$. Show that neither factor is a 7th power in $\mathbb{F}_7[x]$ and that each is relatively prime to its derivative to conclude that both factors are irreducible polynomials, giving the complete factorization of $f(x)$ into irreducible polynomials:

$$f(x) = (x^2 + 3x + 5)(x^3 + 4x^2 + 4x + 4) \in \mathbb{F}_7[x].$$

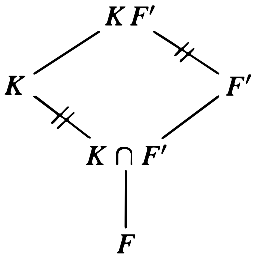
14.4 COMPOSITE EXTENSIONS AND SIMPLE EXTENSIONS

We now consider the effect of taking composites with Galois extensions. The first result states that "sliding up" a Galois extension gives a Galois extension.

Proposition 19. Suppose K/F is a Galois extension and F'/F is any extension. Then KF'/F' is a Galois extension, with Galois group

$$\mathrm{Gal}(KF'/F') \cong \mathrm{Gal}(K/K \cap F')$$

isomorphic to a subgroup of $\mathrm{Gal}(K/F)$. Pictorially,



Proof: If K/F is Galois, then K is the splitting field of some separable polynomial $f(x)$ in $F[x]$. Then KF'/F' is the splitting field of $f(x)$ viewed as a polynomial in

$F'[x]$, hence this extension is Galois. Since K/F is Galois, every embedding of K fixing F is an automorphism of K, so the map

$$\varphi : \mathrm{Gal}(KF'/F') \to \mathrm{Gal}(K/F)$$
$$\sigma \mapsto \sigma|_K$$

defined by restricting an automorphism σ to the subfield K is well defined. It is clearly a homomorphism, with kernel

$$\ker \varphi = \{\sigma \in \mathrm{Gal}(KF'/F') \mid \sigma|_K = 1\}.$$

Since an element in $\mathrm{Gal}(KF'/F')$ is trivial on F', the elements in the kernel are trivial both on K and on F', hence on their composite, so the kernel consists only of the identity automorphism. Hence φ is injective.

Let H denote the image of φ in $\mathrm{Gal}(K/F)$ and let K_H denote the corresponding fixed subfield of K containing F. Since every element in H fixes F', K_H contains $K \cap F'$. On the other hand, the composite $K_H F'$ is fixed by $\mathrm{Gal}(KF'/F')$ (any $\sigma \in \mathrm{Gal}(KF'/F')$ fixes F' and acts on $K_H \subseteq K$ via its restriction $\sigma|_K \in H$, which fixes K_H by definition). By the Fundamental Theorem it follows that $K_H F' = F'$, so that $K_H \subseteq F'$, which gives the reverse inclusion $K_H \subseteq K \cap F'$. Hence $K_H = K \cap F'$, so again by the Fundamental Theorem, $H = \mathrm{Gal}(K/K \cap F')$, completing the proof.

Corollary 20. Suppose K/F is a Galois extension and F'/F is any finite extension. Then

$$[KF' : F] = \frac{[K : F][F' : F]}{[K \cap F' : F]}.$$

Proof: This follows by the proposition from the equality $[KF' : F'] = [K : K \cap F']$ given by the orders of the Galois groups in the proposition.

The example $F = \mathbb{Q}$, $K = \mathbb{Q}(\sqrt[3]{2})$, $F' = \mathbb{Q}(\rho\sqrt[3]{2})$, ρ a primitive 3rd root of unity, shows that the formula of Corollary 20 does not hold in general if neither of the two extensions is Galois.

Proposition 21. Let K_1 and K_2 be Galois extensions of a field F. Then

(1) The intersection $K_1 \cap K_2$ is Galois over F.

(2) The composite $K_1 K_2$ is Galois over F. The Galois group is isomorphic to the subgroup

$$H = \{(\sigma, \tau) \mid \sigma|_{K_1 \cap K_2} = \tau|_{K_1 \cap K_2}\}$$

of the direct product $\mathrm{Gal}(K_1/F) \times \mathrm{Gal}(K_2/F)$ consisting of elements whose restrictions to the intersection $K_1 \cap K_2$ are equal.

$$\begin{tikzcd}
 & K_1K_2 \arrow[dl, dash] \arrow[dr, dash] & \\
K_1 \arrow[dr, dash] & & K_2 \arrow[dl, dash] \\
 & K_1 \cap K_2 \arrow[d, dash] & \\
 & F &
\end{tikzcd}$$

Proof: (1) Suppose $p(x)$ is an irreducible polynomial in $F[x]$ with a root α in $K_1 \cap K_2$. Since $\alpha \in K_1$ and K_1/F is Galois, all the roots of $p(x)$ lie in K_1. Similarly all the roots lie in K_2, hence all the roots of $p(x)$ lie in $K_1 \cap K_2$. It follows easily that $K_1 \cap K_2$ is Galois as in Theorem 13.

(2) If K_1 is the splitting field of the separable polynomial $f_1(x)$ and K_2 is the splitting field of the separable polynomial $f_2(x)$ then the composite is the splitting field for the squarefree part of the polynomial $f_1(x)f_2(x)$, hence is Galois over F.

The map

$$\begin{aligned}\varphi : \mathrm{Gal}(K_1K_2/F) &\to \mathrm{Gal}(K_1/F) \times \mathrm{Gal}(K_2/F)\\ \sigma &\mapsto (\sigma|_{K_1}, \sigma|_{K_2})\end{aligned}$$

is clearly a homomorphism. The kernel consists of the elements σ which are trivial on both K_1 and K_2, hence trivial on the composite, so the map is injective. The image lies in the subgroup H, since

$$(\sigma|_{K_1})|_{K_1\cap K_2} = \sigma|_{K_1\cap K_2} = (\sigma|_{K_2})|_{K_1\cap K_2}.$$

The order of H can be computed by observing that for every $\sigma \in \mathrm{Gal}(K_1/F)$ there are $|\mathrm{Gal}(K_2/K_1 \cap K_2)|$ elements $\tau \in \mathrm{Gal}(K_2/F)$ whose restrictions to $K_1 \cap K_2$ are $\sigma|_{K_1\cap K_2}$. Hence

$$\begin{aligned}|H| &= |\mathrm{Gal}(K_1/F)| \cdot |\mathrm{Gal}(K_2/K_1 \cap K_2)|\\ &= |\mathrm{Gal}(K_1/F)| \frac{|\mathrm{Gal}(K_2/F)|}{|\mathrm{Gal}(K_1 \cap K_2/F)|}.\end{aligned}$$

By Corollary 20 and the diagram above we see that the orders of H and $\mathrm{Gal}(K_1K_2/F)$ are then both equal to

$$[K_1K_2 : F] = \frac{[K_1 : F][K_2 : F]}{[K_1 \cap K_2 : F]}.$$

Hence the image of φ is precisely H, completing the proof.

Corollary 22. Let K_1 and K_2 be Galois extensions of a field F with $K_1 \cap K_2 = F$. Then

$$\mathrm{Gal}(K_1K_2/F) \cong \mathrm{Gal}(K_1/F) \times \mathrm{Gal}(K_2/F).$$

Conversely, if K is Galois over F and $G = \mathrm{Gal}(K/F) = G_1 \times G_2$ is the direct product of two subgroups G_1 and G_2, then K is the composite of two Galois extensions K_1 and K_2 of F with $K_1 \cap K_2 = F$.

Proof: The first part follows immediately from the proposition. For the second, let K_1 be the fixed field of $G_1 \subset G$ and let K_2 be the fixed field of $G_2 \subset G$. Then $K_1 \cap K_2$ is the field corresponding to the subgroup G_1G_2, which is all of G in this case, so $K_1 \cap K_2 = F$. The composite K_1K_2 is the field corresponding to the subgroup $G_1 \cap G_2$, which is the identity here, so $K_1K_2 = K$, completing the proof.

Corollary 23. Let E/F be any finite separable extension. Then E is contained in an extension K which is Galois over F and is minimal in the sense that in a fixed algebraic closure of K any other Galois extension of F containing E contains K.

Proof: There exists a Galois extension of F containing E, for example the composite of the splitting fields of the minimal polynomials for a basis for E over F (which are all separable since E is separable over F). Then the intersection of all the Galois extensions of F containing E is the field K.

Definition. The Galois extension K of F containing E in the previous corollary is called the *Galois closure* of E over F.

It is often simpler to work in a Galois extension (for example in computing degrees as in Corollary 20). The existence of a Galois closure for a separable extension is frequently useful for reducing computations to consideration of Galois extensions.

Recall that an extension K of F is called *simple* if $K = F(\theta)$ for some element θ, in which case θ is called a *primitive element* for K.

Proposition 24. Let K/F be a finite extension. Then $K = F(\theta)$ if and only if there exist only finitely many subfields of K containing F.

Proof: Suppose first that $K = F(\theta)$ is simple. Let E be a subfield of K containing F: $F \subseteq E \subseteq K$. Let $f(x) \in F[x]$ be the minimal polynomial for θ over F and let $g(x) \in E[x]$ be the minimal polynomial for θ over E. Then $g(x)$ divides $f(x)$ in $E[x]$. Let E' be the field generated over F by the coefficients of $g(x)$. Then $E' \subseteq E$ and clearly the minimal polynomial for θ over E' is still $g(x)$. But then

$$[K : E] = \deg g(x) = [K : E']$$

implies that $E = E'$. It follows that the subfields of K containing F are the subfields generated by the coefficients of the monic factors of $f(x)$, hence there are finitely many such subfields.

Suppose conversely that there are finitely many subfields of K containing F. If F is a finite field, then we have already seen that K is a simple extension (Proposition 17). Hence we may suppose F is infinite. It clearly suffices to show that $F(\alpha, \beta)$ is generated by a single element since K is finitely generated over F. Consider the subfields

$$F(\alpha + c\beta), \qquad c \in F.$$

Then since there are infinitely many choices for $c \in F$ and only finitely many such subfields, there exist c, c' in F, $c \neq c'$, with

$$F(\alpha + c\beta) = F(\alpha + c'\beta).$$

Then $\alpha + c\beta$ and $\alpha + c'\beta$ both lie in $F(\alpha + c\beta)$, and taking their difference shows that $(c - c')\beta \in F(\alpha + c\beta)$ Hence $\beta \in F(\alpha + c\beta)$ and then also $\alpha \in F(\alpha + c\beta)$. Therefore $F(\alpha, \beta) \subseteq F(\alpha + c\beta)$ and since the reverse inclusion is obvious, we have

$$F(\alpha, \beta) = F(\alpha + c\beta),$$

completing the proof.

Theorem 25. *(The Primitive Element Theorem)* If K/F is finite and separable, then K/F is simple. In particular, any finite extension of fields of characteristic 0 is simple.

Proof: Let L be the Galois closure of K over F. Then any subfield of K containing F corresponds to a subgroup of the Galois group $\text{Gal}(L/F)$ by the Fundamental Theorem. Since there are only finitely many such subgroups, the previous proposition shows that K/F is simple. The last statement follows since any finite extension of fields in characteristic 0 is separable.

As the proof of the proposition indicates, a primitive element for an extension can be obtained as a simple linear combination of the generators for the extension. In the case of Galois extensions it is only necessary to determine a linear combination which is not fixed by any nontrivial element of the Galois group since then by the Fundamental Theorem this linear combination could not lie in any proper subfield.

Examples

(1) The element $\sqrt{2}+\sqrt{3}$ generates the field $\mathbb{Q}(\sqrt{2},\sqrt{3})$ as we have already seen (it is not fixed by any of the four Galois automorphisms of this field).

(2) The field $\overline{\mathbb{F}_p}(x, y)$ of rational functions in the variables x and y over the algebraic closure $\overline{\mathbb{F}_p}$ of $\mathbb{F}_p$ is not a simple extension of the subfield $F = \overline{\mathbb{F}_p}(x^p, y^p)$. It is easy to see that

$$[\overline{\mathbb{F}_p}(x, y) : \overline{\mathbb{F}_p}(x^p, y^p)] = p^2$$

and that the subfields

$$F(x + cy), \qquad c \in \overline{\mathbb{F}_p}$$

are all of degree p over $\overline{\mathbb{F}_p}(x^p, y^p)$ (note that $(x+cy)^p = x^p + c^p y^p \in \overline{\mathbb{F}_p}(x^p, y^p)$). If any two of these subfields were equal, then just as in the proof of Proposition 24 we would have

$$\overline{\mathbb{F}_p}(x, y) = F(x + cy)$$

which is impossible by degree considerations. Hence there are infinitely many such subfields and the extension cannot be simple.

EXERCISES

1. Determine the Galois closure of the field $\mathbb{Q}(\sqrt{1+\sqrt{2}})$ over $\mathbb{Q}$.
2. Find a primitive generator for $\mathbb{Q}(\sqrt{2},\sqrt{3},\sqrt{5})$ over $\mathbb{Q}$.
3. Let F be a field contained in the ring of $n \times n$ matrices over $\mathbb{Q}$. Prove that $[F : \mathbb{Q}] \le n$. (Note that, by Exercise 19 of Section 13.2, the ring of $n \times n$ matrices over $\mathbb{Q}$ does contain fields of degree n over $\mathbb{Q}$.)
4. Let $f(x) \in F[x]$ be an irreducible polynomial of degree n over the field F, let L be the splitting field of $f(x)$ over F and let α be a root of $f(x)$ in L. If K is any Galois extension of F, show that the polynomial $f(x)$ splits into a product of m irreducible polynomials each of degree d over K, where $d = [K(\alpha) : K] = [(L \cap K)(\alpha) : L \cap K]$ and $m = n/d = [F(\alpha) \cap K : F]$. [Show first that the factorization of $f(x)$ over K is the same as its factorization over $L \cap K$. Then if H is the subgroup of the Galois group of L

over F corresponding to $L \cap K$ the factors of $f(x)$ over $L \cap K$ correspond to the orbits of H on the roots of $f(x)$. Use Exercise 9 of Section 4.1.]

5. Let p be a prime and let F be a field. Let K be a Galois extension of F whose Galois group is a p-group (i.e., the degree $[K : F]$ is a power of p). Such an extension is called a *p-extension* (note that p-extensions are Galois by definition).

  (a) Let L be a p-extension of K. Prove that the Galois closure of L over F is a p-extension of F.

  (b) Give an example to show that (a) need not hold if $[K : F]$ is a power of p but K/F is not Galois.

6. Prove that $\mathbb{F}_p(x, y)/\mathbb{F}_p(x^p, y^p)$ is not a simple extension by explicitly exhibiting an infinite number of intermediate subfields.

7. Let $F \subseteq K \subseteq L$ and let $\theta \in L$ with $p(x) = m_{\theta,F}(x)$. Prove that $K \otimes_F F(\theta) \cong K[x]/(p(x))$ as K-algebras.

8. Let K_1 and K_2 be two algebraic extensions of a field F contained in the field L of characteristic zero. Prove that the F-algebra $K_1 \otimes_F K_2$ has no nonzero nilpotent elements. [Use the preceding exercise.]

14.5 CYCLOTOMIC EXTENSIONS AND ABELIAN EXTENSIONS OVER $\mathbb{Q}$

We have already determined that the cyclotomic field $\mathbb{Q}(\zeta_n)$ of n^{th} roots of unity is a Galois extension of $\mathbb{Q}$ of degree $\varphi(n)$ where φ denotes the Euler φ-function. Any automorphism of this field is uniquely determined by its action on the primitive n^{th} root of unity ζ_n. This element must be mapped to another primitive n^{th} root of unity (recall these are the roots of the irreducible cyclotomic polynomial $\Phi_n(x)$). Hence $\sigma(\zeta_n) = \zeta_n^a$ for some integer a, $1 \leq a < n$, relatively prime to n. Since there are precisely $\varphi(n)$ such integers a it follows that in fact each of these maps is indeed an automorphism of $\mathbb{Q}(\zeta_n)$. Note also that we can define σ_a for any integer a relatively prime to n by the same formula and that σ_a depends only on the residue class of a modulo n.

Theorem 26. The Galois group of the cyclotomic field $\mathbb{Q}(\zeta_n)$ of n^{th} roots of unity is isomorphic to the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$. The isomorphism is given explicitly by the map

$$\begin{aligned} (\mathbb{Z}/n\mathbb{Z})^\times &\xrightarrow{\sim} \text{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q}) \\ a \pmod{n} &\longmapsto \sigma_a \end{aligned}$$

where σ_a is the automorphism defined by

$$\sigma_a(\zeta_n) = \zeta_n^a.$$

Proof: The discussion above shows that σ_a is an automorphism for any $a \pmod{n}$, so the map above is well defined. It is a homomorphism since

$$\begin{aligned} (\sigma_a\sigma_b)(\zeta_n) = \sigma_a(\zeta_n^b) &= (\zeta_n^b)^a \\ &= \zeta_n^{ab} \end{aligned}$$

which shows that $\sigma_a\sigma_b = \sigma_{ab}$. The map is bijective by the discussion above since we know that every Galois automorphism is of the form σ_a for a uniquely defined a (mod n). Hence the map is an isomorphism.

Examples

(1) The field $\mathbb{Q}(\zeta_5)$ is Galois over $\mathbb{Q}$ with Galois group $(\mathbb{Z}/5\mathbb{Z})^\times \cong \mathbb{Z}/4\mathbb{Z}$. This is our first example of a Galois extension of $\mathbb{Q}$ of degree 4 with a *cyclic* Galois group. The elements of the Galois group are $\{\sigma_1 = 1, \sigma_2, \sigma_3, \sigma_4\}$ in the notation above. A generator for this cyclic group is $\sigma_2 : \zeta_5 \mapsto \zeta_5^2$ (since 2 has order 4 in $(\mathbb{Z}/5\mathbb{Z})^\times$).

There is precisely one nontrivial subfield, a quadratic extension of $\mathbb{Q}$, the fixed field of the subgroup $\{1, \sigma_4 = \sigma_{-1}\}$. An element in this subfield is given by

$$\alpha = \zeta_5 + \sigma_{-1}\zeta_5 = \zeta_5 + \zeta_5^{-1}$$

since this element is clearly fixed by σ_{-1}. The element ζ_5 satisfies

$$\zeta_5^4 + \zeta_5^3 + \zeta_5^2 + \zeta_5 + 1 = 0.$$

Notice then that

$$\begin{aligned}\alpha^2 + \alpha - 1 &= (\zeta_5^2 + 2 + \zeta_5^{-2}) + (\zeta_5 + \zeta_5^{-1}) - 1\\ &= \zeta_5^2 + 2 + \zeta_5^3 + \zeta_5 + \zeta_5^4 - 1 = 0.\end{aligned}$$

Solving explicitly for α we see that the quadratic extension of $\mathbb{Q}$ generated by α is $\mathbb{Q}(\sqrt{5})$:

$$\mathbb{Q}(\zeta_5 + \zeta_5^{-1}) = \mathbb{Q}(\sqrt{5}).$$

It can be shown in general (this is not completely trivial) that for p an odd prime the field $\mathbb{Q}(\zeta_p)$ contains the quadratic field $\mathbb{Q}(\sqrt{\pm p})$, where the $+$ sign is correct if $p \equiv 1 \bmod 4$ and the $-$ sign is correct if $p \equiv 3 \bmod 4$ (cf. Exercise 11 in Section 7).

(2) $\mathbb{Q}(\zeta_{13})$,For p an odd prime we can construct a primitive element for any of the subfields of $\mathbb{Q}(\zeta_p)$ as in the previous example. A basis for $\mathbb{Q}(\zeta_p)$ over $\mathbb{Q}$ is given by

$$1, \zeta_p, \zeta_p^2, \ldots, \zeta_p^{p-2}.$$

Since

$$\zeta_p^{p-1} + \zeta_p^{p-2} + \cdots + \zeta_p + 1 = 0$$

we see that also the elements

$$\zeta_p, \zeta_p^2, \ldots, \zeta_p^{p-2}, \zeta_p^{p-1}$$

form a basis. The reason for choosing this basis is that any σ in the Galois group $\mathrm{Gal}(\mathbb{Q}(\zeta_p)/\mathbb{Q})$ simply *permutes* these basis elements since these are precisely the primitive p^{th} roots of unity. Note that it is at this point that we need p to be a prime — in general the primitive n^{th} roots of unity do not give a basis for the cyclotomic field of n^{th} roots of unity over $\mathbb{Q}$ (for example, the primitive 4th roots of unity, $\pm i$, are not linearly independent).

Let H be any subgroup of the Galois group of $\mathbb{Q}(\zeta_p)$ over $\mathbb{Q}$ and let

$$\alpha_H = \sum_{\sigma \in H} \sigma\zeta_p, \tag{14.10}$$

the sum of the conjugates of ζ_p by the elements in H. For any $\tau \in H$, the elements $\tau\sigma$ run over the elements of H as σ runs over the elements of H. It follows that $\tau\alpha = \alpha$, so that α lies in the fixed field for H. If now τ is *not* an element of H, then $\tau\alpha$ is the sum of basis elements (recall that any automorphism permutes the basis elements here), one of which is $\tau(\zeta_p)$. If we had $\tau\alpha = \alpha$ then since these elements are a basis, we must have $\tau(\zeta_p) = \sigma(\zeta_p)$ for one of the terms $\sigma\zeta_p$ in (10). But this implies $\tau\sigma^{-1} = 1$ since this automorphism is the identity on ζ_p. Then $\tau = \sigma \in H$, a contradiction. This shows that α is not fixed by any automorphism not contained in H, so that $\mathbb{Q}(\alpha)$ is precisely the fixed field of H.

For a specific example, consider the subfields of $\mathbb{Q}(\zeta_{13})$, which correspond to the subgroups of $(\mathbb{Z}/13\mathbb{Z})^\times \cong \mathbb{Z}/12\mathbb{Z}$. A generator for this cyclic group is the automorphism $\sigma = \sigma_2$ which maps ζ_{13} to ζ_{13}^2. The nontrivial subgroups correspond to the nontrivial divisors of 12, hence are of orders 2, 3, 4, and 6 with generators σ^6, σ^4, σ^3 and σ^2, respectively. The corresponding fixed fields will be of degrees 6, 4, 3 and 2 over $\mathbb{Q}$, respectively. Generators are given by ($\zeta = \zeta_{13}$)

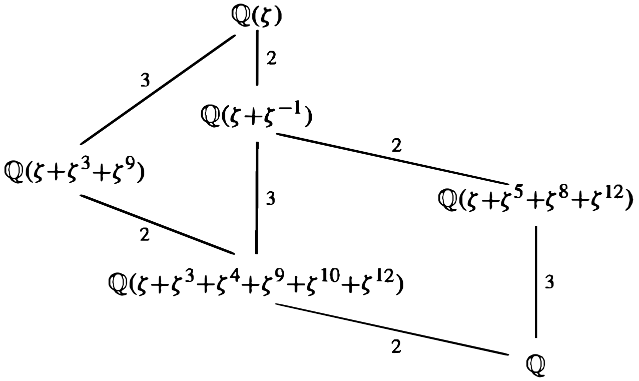
$$\zeta + \sigma^6\zeta = \zeta + \zeta^{2^6} = \zeta + \zeta^{-1}$$

$$\zeta + \sigma^4\zeta + \sigma^8\zeta = \zeta + \zeta^{2^4} + \zeta^{2^8} = \zeta + \zeta^3 + \zeta^9$$

$$\zeta + \sigma^3\zeta + \sigma^6\zeta + \sigma^9\zeta = \zeta + \zeta^8 + \zeta^{12} + \zeta^5$$

$$\zeta + \sigma^2\zeta + \sigma^4\zeta + \sigma^6\zeta + \sigma^8\zeta + \sigma^{10}\zeta = \zeta + \zeta^4 + \zeta^3 + \zeta^{12} + \zeta^9 + \zeta^{10}.$$

The lattice of subfields for this extension is the following:



$\mathbb{Q}(\zeta)$
2
3
$\mathbb{Q}(\zeta+\zeta^{-1})$
2
$\mathbb{Q}(\zeta+\zeta^3+\zeta^9)$
3
$\mathbb{Q}(\zeta+\zeta^5+\zeta^8+\zeta^{12})$
2
$\mathbb{Q}(\zeta+\zeta^3+\zeta^4+\zeta^9+\zeta^{10}+\zeta^{12})$
3
2
$\mathbb{Q}$

The elements constructed in equation (10) and their conjugates are called the *periods* of ζ and are useful in the study of the arithmetic of the cyclotomic fields. The study of their combinatorial properties is referred to as *cyclotomy*.

Suppose that $n = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$ is the decomposition of n into distinct prime powers. Since $\zeta_n^{p_2^{a_2}\cdots p_k^{a_k}}$ is a primitive $p_1^{a_1}$-th root of unity, the field $K_1 = \mathbb{Q}(\zeta_{p_1^{a_1}})$ is a subfield of $\mathbb{Q}(\zeta_n)$. Similarly, each of the fields $K_i = \mathbb{Q}(\zeta_{p_i^{a_i}})$, $i = 1, 2, \dots, k$ is a subfield of $\mathbb{Q}(\zeta_n)$. The composite of the fields contains the product $\zeta_{p_1^{a_1}}\zeta_{p_2^{a_2}} \cdots \zeta_{p_k^{a_k}}$, which is a primitive n^{th} root of unity, hence the composite field is $\mathbb{Q}(\zeta_n)$. Since the extension degrees $[K_i : \mathbb{Q}]$ equal $\varphi(p_i^{a_i})$, $i = 1, 2, \dots, k$ and $\varphi(n) = \varphi(p_1^{a_1})\varphi(p_2^{a_2}) \cdots \varphi(p_k^{a_k})$, the degree of the composite of the fields K_i is precisely the product of the degrees of the K_i. It follows from Proposition 21 (and a simple induction from the two fields considered in the proposition to the k fields here) that the intersection of all these fields is precisely $\mathbb{Q}$. Then Corollary 22 shows that the Galois group for $\mathbb{Q}(\zeta_n)$ is the direct product of the Galois groups over $\mathbb{Q}$ for the subfields K_i. We summarize this as the following corollary.

Corollary 27. Let $n = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$ be the decomposition of the positive integer n into distinct prime powers. Then the cyclotomic fields $\mathbb{Q}(\zeta_{p_i^{a_i}})$, $i = 1, 2, \ldots, k$ intersect only in the field $\mathbb{Q}$ and their composite is the cyclotomic field $\mathbb{Q}(\zeta_n)$. We have

$$\mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q}) \cong \mathrm{Gal}(\mathbb{Q}(\zeta_{p_1^{a_1}})/\mathbb{Q}) \times \mathrm{Gal}(\mathbb{Q}(\zeta_{p_2^{a_2}})/\mathbb{Q}) \times \cdots \times \mathrm{Gal}(\mathbb{Q}(\zeta_{p_k^{a_k}})/\mathbb{Q})$$

which under the isomorphism in Theorem 26 is the Chinese Remainder Theorem:

$$(\mathbb{Z}/n\mathbb{Z})^\times \cong (\mathbb{Z}/p_1^{a_1}\mathbb{Z})^\times \times (\mathbb{Z}/p_2^{a_2}\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/p_k^{a_k}\mathbb{Z})^\times.$$

Proof: The only statement which has not been proved is the identification of the isomorphism of Galois groups with the statement of the Chinese Remainder Theorem on the group $(\mathbb{Z}/n\mathbb{Z})^\times$, which is quite simple and is left for the exercises.

By Theorem 26 the Galois group of $\mathbb{Q}(\zeta_n)/\mathbb{Q}$ is in particular an abelian group.

Definition. The extension K/F is called an *abelian* extension if K/F is Galois and $\mathrm{Gal}(K/F)$ is an abelian group.

Since all the subgroups and quotient groups of abelian groups are abelian, we see by the Fundamental Theorem of Galois Theory that every subfield containing F of an abelian extension of F is again an abelian extension of F. By the results on composites of extensions in the last section, we also see that the composite of abelian extensions is again an abelian extension (since the Galois group of the composite is isomorphic to a subgroup of the direct product of the Galois groups, hence is abelian).

It is an open problem to determine which groups arise as the Galois groups of Galois extensions of $\mathbb{Q}$. Using the results above we can see that every *abelian* group appears as the Galois group of some extension of $\mathbb{Q}$, in fact as the Galois group of some subfield of a cyclotomic field.

Let $n = p_1 p_2 \cdots p_k$ be the product of distinct primes. Then by the Chinese Remainder Theorem

$$\begin{aligned}(\mathbb{Z}/n\mathbb{Z})^\times &\cong (\mathbb{Z}/p_1\mathbb{Z})^\times \times (\mathbb{Z}/p_2\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/p_k\mathbb{Z})^\times \\ &\cong Z_{p_1-1} \times Z_{p_2-1} \times \cdots \times Z_{p_k-1}. \end{aligned} \qquad (14.11)$$

Now, suppose G is any finite abelian group. By the Fundamental Theorem for Abelian Groups,

$$G \cong Z_{n_1} \times Z_{n_2} \times \cdots \times Z_{n_k}$$

for some integers $n_1, n_2, \ldots, n_k$. We take as known that given any integer m there are infinitely many primes p with $p \equiv 1 \bmod m$ (see the exercises following Section 13.6

for one proof using cyclotomic polynomials). Given this result, choose distinct primes $p_1, p_2, \ldots, p_k$ such that

$$\begin{aligned} p_1 &\equiv 1 \bmod n_1 \\ p_2 &\equiv 1 \bmod n_2 \\ &\vdots \\ p_k &\equiv 1 \bmod n_k \end{aligned}$$

and let $n = p_1 p_2 \cdots p_k$ as above.

By construction, n_i divides $p_i - 1$ for $i = 1, 2, \ldots, k$, so the group $Z_{p_i - 1}$ has a subgroup H_i of order $\dfrac{p_i - 1}{n_i}$ for $i = 1, 2, \ldots, k$, and the quotient by this subgroup is cyclic of order n_i. Hence the quotient of $(\mathbb{Z}/n\mathbb{Z})^\times$ in equation (11) by $H_1 \times H_2 \times \cdots \times H_k$ is isomorphic to the group G.

By Theorem 26 and the Fundamental Theorem of Galois Theory, we see that there is a subfield of $\mathbb{Q}(\zeta_{p_1 p_2 \cdots p_k})$ which is Galois over $\mathbb{Q}$ with G as Galois group. We summarize this in the following corollary.

Corollary 28. Let G be any finite abelian group. Then there is a subfield K of a cyclotomic field with $\mathrm{Gal}(K/\mathbb{Q}) \cong G$.

There is a converse to this result (whose proof is beyond our scope), the celebrated Kronecker–Weber Theorem:

Theorem *(Kronecker–Weber)* Let K be a finite abelian extension of $\mathbb{Q}$. Then K is contained in a cyclotomic extension of $\mathbb{Q}$.

The abelian extensions of $\mathbb{Q}$ are the "easiest" Galois extensions (at least in so far as the structure of their Galois groups is concerned) and the previous result shows they can be classified by the cyclotomic extensions of $\mathbb{Q}$. For other finite extensions of $\mathbb{Q}$ as base field, it is more difficult to describe the abelian extensions. The study of the abelian extensions of an arbitrary finite extension F of $\mathbb{Q}$ is referred to as *class field theory*. There is a classification of the abelian extensions of F by invariants associated to F which greatly generalizes the results on cyclotomic fields over $\mathbb{Q}$. In general, however, the construction of abelian extensions is not nearly as explicit as in the case of the cyclotomic fields. One case where such a description is possible is for the abelian extensions of an imaginary quadratic field ($\mathbb{Q}(\sqrt{-D})$ for D positive), where the abelian extensions can be constructed by adjoining values of certain elliptic functions (this is the analogue of adjoining the roots of unity, which are the values of the exponential function e^x for certain x). The study of the arithmetic of such abelian extensions and the search for similar results for non-abelian extensions are rich and fascinating areas of current mathematical research.

We end our discussion of the cyclotomic fields with the problem of the constructibility of the regular n-gon by straightedge and compass.

Recall (cf. Section 13.3) that an element α is constructible over $\mathbb{Q}$ if and only if the field $\mathbb{Q}(\alpha)$ is contained in a field K obtained by a series of quadratic extensions:

$$\mathbb{Q} = K_0 \subset K_1 \subset \cdots \subset K_i \subset K_{i+1} \subset \cdots \subset K_m = K \tag{14.12}$$

with

$$[K_{i+1} : K_i] = 2, \qquad i = 0, 1, \ldots, m-1.$$

The construction of the regular n-gon in $\mathbb{R}^2$ is evidently equivalent to the construction of the n^{th} roots of unity, since the n^{th} roots of unity form the vertices of a regular n-gon on the unit circle in $\mathbb{C}$ with one vertex at the point 1.

The construction of ζ_n is equivalent to the constructibility of the first coordinate x in $\mathbb{R}^2$ of ζ_n, namely the real part of ζ_n. Since the complex conjugate of ζ_n is just ζ_n^{-1}, the real part of ζ_n is $x = \frac{1}{2}(\zeta_n + \zeta_n^{-1})$. Note that ζ_n satisfies the quadratic equation $\zeta_n^2 - 2x\zeta_n + 1 = 0$ over $\mathbb{Q}(x)$. Since $\mathbb{Q}(x)$ consists only of real numbers, it follows that $[\mathbb{Q}(\zeta_n) : \mathbb{Q}(x)] = 2$, so that $\mathbb{Q}(x)$ is an extension of degree $\varphi(n)/2$ of $\mathbb{Q}$.

It follows that if the regular n-gon can be constructed by straightedge and compass then $\varphi(n)$ must be a power of 2. Conversely, if $\varphi(n) = 2^m$ is a power of 2, then the Galois group $\mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$ is an abelian group whose order is a power of 2, so the same is true for the Galois group $\mathrm{Gal}(\mathbb{Q}(x)/\mathbb{Q})$. It is easy to see by the Fundamental Theorem for Abelian Groups that an abelian group G of order 2^m has a chain of subgroups

$$G = G_m > G_{m-1} > \cdots > G_{i+1} > G_i > \cdots > G_0 = 1$$

with

$$[G_{i+1} : G_i] = 2, \qquad i = 0, 1, 2, \ldots, m-1.$$

Applying this to the group $G = \mathrm{Gal}(\mathbb{Q}(x)/\mathbb{Q})$ and taking the fixed fields for the subgroups G_i, $i = 0, 1, \ldots, m-1$, we obtain (by the Fundamental Theorem of Galois Theory) a sequence of quadratic extensions as in (12) above.

We conclude that the regular n-gon can be constructed by straightedge and compass if and only if $\varphi(n)$ is a power of 2. Decomposing n into prime powers to compute $\varphi(n)$ we see that this means $n = 2^k p_1 \cdots p_r$ is the product of a power of 2 and distinct odd primes p_i where $p_i - 1$ is a power of 2. It is an elementary exercise to see that a prime p with $p - 1$ a power of 2 must be of the form

$$p = 2^{2^s} + 1$$

for some integer s. Such primes are called *Fermat primes*. The first few are

$$3 = 2^1 + 1$$
$$5 = 2^2 + 1$$
$$17 = 2^4 + 1$$
$$257 = 2^8 + 1$$
$$65537 = 2^{16} + 1$$

(but $2^{32} + 1$ is not a prime, being divisible by 641). It is not known if there are infinitely many Fermat primes. We summarize this in the following proposition.

Proposition 29. The regular n-gon can be constructed by straightedge and compass if and only if $n = 2^k p_1 \cdots p_r$ is the product of a power of 2 and distinct Fermat primes.

The proof above actually indicates a procedure for constructing the regular n-gon as a succession of square roots. For example, the construction of the regular 17-gon (solved by Gauss in 1796 at age 19) requires the construction of the subfields of degrees 2, 4, 8 and 16 in $\mathbb{Q}(\zeta_{17})$. These subfields can be constructed by forming the *periods* of ζ_{17} as in the example of the 13th roots of unity above. In this case, the fact that $\mathbb{Q}(\zeta_{17})$ is obtained by a series of quadratic extensions reflects itself in the fact that the periods can be "halved" successively (i.e., if $H_1 < H_2$ are subgroups with $[H_2 : H_1] = 2$ then the periods for H_1 satisfy a quadratic equation whose coefficients involve the periods for H_2). For example, the periods for the subgroup of index 2 (generated by σ_2) in the Galois group are ($\zeta = \zeta_{17}$)

$$\eta_1 = \zeta + \zeta^2 + \zeta^4 + \zeta^8 + \zeta^9 + \zeta^{13} + \zeta^{15} + \zeta^{16}$$
$$\eta_2 = \zeta^3 + \zeta^5 + \zeta^6 + \zeta^7 + \zeta^{10} + \zeta^{11} + \zeta^{12} + \zeta^{14}$$

which "halve" the period for the full Galois group and which satisfy

$$\eta_1 + \eta_2 = -1$$

(from the minimal polynomial satisfied by ζ_{17}) and

$$\eta_1\eta_2 = -4$$

(which requires computation — we know that it must be rational by Galois Theory, since this product is fixed by all the elements of the Galois group). Hence these two periods are the roots of the quadratic equation

$$x^2 + x - 4 = 0$$

which we can solve explicitly. In a similar way, the periods for the subgroup of index 4 (generated by σ_4) naturally halve these periods, so are quadratic over these, etc. In this way one can determine ζ_{17} explicitly in terms of iterated square roots. For example, one finds that $8(\zeta + \zeta^{-1}) = 16\cos(\frac{2\pi}{17})$ (which is enough to construct the regular 17-gon) is given explicitly by

$$-1+\sqrt{17}+\sqrt{2(17-\sqrt{17})}+2\sqrt{17+3\sqrt{17}-\sqrt{2(17-\sqrt{17})}-2\sqrt{2(17+\sqrt{17})}}\,.$$

A relatively simple construction of the regular 17-gon (shown to us by J.H. Conway) is indicated in the exercises.

While we have seen that it is not possible to solve for ζ_n using only successive square roots in general, by definition it is possible to obtain ζ_n by successive extraction of higher roots (namely, taking an n^{th} root of 1). This is not the case for solutions of general equations of degree n, where one cannot generally determine solutions by radicals, as we shall see in the next sections.

EXERCISES

1. Determine the minimal polynomials satisfied by the primitive generators given in the text for the subfields of $\mathbb{Q}(\zeta_{13})$.

2. Determine the subfields of $\mathbb{Q}(\zeta_8)$ generated by the periods of ζ_8 and in particular show that not every subfield has such a period as primitive element.

3. Determine the quadratic equation satisfied by the period $\alpha = \zeta_5 + \zeta_5^{-1}$ of the 5th root of unity ζ_5. Determine the quadratic equation satisfied by ζ_5 over $\mathbb{Q}(\alpha)$ and use this to explicitly solve for the 5th root of unity.

4. Let $\sigma_a \in \mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$ denote the automorphism of the cyclotomic field of n^{th} roots of unity which maps ζ_n to ζ_n^a where a is relatively prime to n and ζ_n is a primitive n^{th} root of unity. Show that $\sigma_a(\zeta) = \zeta^a$ for *every* n^{th} root of unity.

5. Let p be a prime and let $\epsilon_1, \epsilon_2, \dots, \epsilon_{p-1}$ denote the primitive p^{th} roots of unity. Set $p_n = \epsilon_1^n + \epsilon_2^n + \cdots + \epsilon_{p-1}^n$, the sum of the n^{th} powers of the ϵ_i. Prove that $p_n = -1$ if p does not divide n and that $p_n = p - 1$ if p does divide n. [One approach: $p_1 = -1$ from $\Phi_p(x)$; show that p_n is a Galois conjugate of p_1 for p not dividing n, hence is also -1.]

6. Let ζ_n denote a primitive n^{th} root of unity and let $K = \mathbb{Q}(\zeta_n)$ be the associated cyclotomic field. Let a denote the trace of ζ_n from K to $\mathbb{Q}$ (cf. Exercise 18 of Section 2). Prove that $a = 1$ if $n = 1$, $a = 0$ if n is divisible by the square of a prime, and $a = (-1)^r$ if n is the product of r distinct primes.

7. Show that complex conjugation restricts to the automorphism $\sigma_{-1} \in \mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$ of the cyclotomic field of n^{th} roots of unity. Show that the field $K^+ = \mathbb{Q}(\zeta_n + \zeta_n^{-1})$ is the subfield of real elements in $K = \mathbb{Q}(\zeta_n)$, called the *maximal real subfield of* K.

8. Let $K_n = \mathbb{Q}(\zeta_{2^{n+2}})$ be the cyclotomic field of 2^{n+2}-th roots of unity, $n \geq 0$. Set $\alpha_n = \zeta_{2^{n+2}} + \zeta_{2^{n+2}}^{-1}$ and $K_n^+ = \mathbb{Q}(\alpha_n)$, the maximal real subfield of K_n.

(a) Show that for all $n \geq 0$, $[K_n : \mathbb{Q}] = 2^{n+1}$, $[K_n : K_n^+] = 2$, $[K_n^+ : \mathbb{Q}] = 2^n$, and $[K_{n+1}^+ : K_n^+] = 2$.

(b) Determine the quadratic equation satisfied by $\zeta_{2^{n+2}}$ over K_n^+ in terms of α_n.

(c) Show that for $n \geq 0$, $\alpha_{n+1}^2 = 2 + \alpha_n$ and hence show that

$$\alpha_n = \sqrt{2 + \sqrt{2 + \sqrt{\cdots + \sqrt{2}}}} \qquad (n \text{ times}),$$

giving an explicit formula for the (constructible) 2^{n+2}-th roots of unity.

9. Notation as in the previous exercise.

(a) Prove that K_n^+ is a cyclic extension of $\mathbb{Q}$ of degree 2^n. [Use an explicit isomorphism $(\mathbb{Z}/2^{n+2}\mathbb{Z})^\times \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2^n\mathbb{Z}$ as abelian groups (i.e., $(\mathbb{Z}/2^{n+2}\mathbb{Z})^\times$ is isomorphic to a cyclic group of order 2 and a cyclic group of order 2^n — cf. Exercises 22 and 23 of Section 2.3]

(b) Prove that K_n is a biquadratic extension of K_{n-1}^+ and that two of the three intermediate subfields are K_n^+ and K_{n-1}. Prove that the remaining field intermediate between K_{n-1}^+ and K_n is a cyclic extension of $\mathbb{Q}$ of degree 2^n.

10. Prove that $\mathbb{Q}(\sqrt[3]{2})$ is not a subfield of any cyclotomic field over $\mathbb{Q}$.

11. Prove that the primitive n^{th} roots of unity form a basis over $\mathbb{Q}$ for the cyclotomic field of n^{th} roots of unity if and only if n is squarefree (i.e., n is not divisible by the square of any prime).

12. Let σ_p denote the Frobenius automorphism $x \mapsto x^p$ of the finite field $\mathbb{F}_q$ of $q = p^n$ elements. Viewing $\mathbb{F}_q$ as a vector space V of dimension n over $\mathbb{F}_p$ we can consider σ_p as a linear transformation of V to V. Determine the characteristic polynomial of σ_p and prove that the linear transformation σ_p is diagonalizable over $\mathbb{F}_p$ if and only if n divides $p-1$, and is diagonalizable over the algebraic closure of $\mathbb{F}_p$ if and only if $(n, p) = 1$.

13. Let $n = p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}$ be the prime factorization of n and let ζ_n be a primitive n^{th} root of unity. For each $i = 1, 2, \dots, k$ define d_i by $n = p_i^{a_i} d_i$ and let $\zeta_{p_i^{a_i}} = \zeta_n^{d_i}$, so that $\zeta_{p_i^{a_i}}$ is a particular primitive $p_i^{a_i}$-th root of unity. Let $\sigma_a \in \operatorname{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$ be the automorphism mapping ζ_n to ζ_n^a for a relatively prime to n.

(a) Prove that for $i = 1, 2, \dots, k$, σ_a maps $\zeta_{p_i^{a_i}}$ to $\zeta_{p_i^{a_i}}^a$ and gives an automorphism of $\mathbb{Q}(\zeta_{p_i^{a_i}})/\mathbb{Q})$ which depends only on $a \pmod{p_i^{a_i}}$, which we may denote $\sigma_{a \,(\mathrm{mod}\, p_i^{a_i})}$.

(b) Prove that the map $\sigma_a \mapsto (\sigma_{a \,(\mathrm{mod}\, p_1^{a_1})}, \dots, \sigma_{a \,(\mathrm{mod}\, p_k^{a_k})})$ is the isomorphism of Corollary 27 corresponding to the Chinese Remainder Theorem for $(\mathbb{Z}/n\mathbb{Z})^\times$.

The following Exercises 14 to 18 determine the periods associated to a primitive 17th root of unity and provide a proof for the simple geometric construction indicated in Exercise 17 for the regular 17-gon. Let $\zeta = \zeta_{17} = \cos \dfrac{2\pi}{17} + i \sin \dfrac{2\pi}{17}$ be a fixed primitive 17th root of unity in $\mathbb{C}$.

14. Define the *periods* of ζ as follows:

$$\begin{array}{ll}
\eta_1 = \zeta+\zeta^2+\zeta^4+\zeta^8+\zeta^9+\zeta^{13}+\zeta^{15}+\zeta^{16} & \eta_3' = \zeta^6+\zeta^7+\zeta^{10}+\zeta^{11} \\
\eta_2 = \zeta^3+\zeta^5+\zeta^6+\zeta^7+\zeta^{10}+\zeta^{11}+\zeta^{12}+\zeta^{14} & \eta_4' = \zeta^3+\zeta^5+\zeta^{12}+\zeta^{14} \\
\eta_1' = \zeta+\zeta^4+\zeta^{13}+\zeta^{16} & \eta_1'' = \zeta+\zeta^{16} \\
\eta_2' = \zeta^2+\zeta^8+\zeta^9+\zeta^{15} & \eta_2'' = \zeta^4+\zeta^{13}.
\end{array}$$

(a) Show that all of these periods are real numbers and that $\eta_1'' = 2\cos \dfrac{2\pi}{17}$. Show that as real numbers these periods are approximately

$$\begin{array}{llll}
\eta_1 \sim 1.562 & \eta_1' \sim 2.049 & \eta_3' \sim -2.906 & \eta_1'' \sim 1.865 \\
\eta_2 \sim -2.562 & \eta_2' \sim -0.488 & \eta_4' \sim 0.344 & \eta_2'' \sim 0.185.
\end{array}$$

(b) Prove that η_1 and η_2 are roots of the equation $x^2 + x - 4 = 0$.

(c) Prove that η_1' and η_2' are roots of the equation $x^2 - \eta_1 x - 1 = 0$ and that η_3' and η_4' are roots of the equation $x^2 - \eta_2 x - 1 = 0$.

(d) Prove that η_1'' and η_2'' are roots of the equation $x^2 - \eta_1' x + \eta_4' = 0$.

15. Prove that if $\tan 2\theta = a$ $(0 < 2\theta < \dfrac{\pi}{2})$ then $\tan\theta$ satisfies the equation $x^2 - \dfrac{2}{a}x - 1 = 0$.

16. Let C be the circle in $\mathbb{R}^2$ having the points (h, k) and $(0, 1)$ as a diameter. Prove that this circle intersects the x-axis if and only if $h^2 - 4k \geq 0$ and in this case the two intercepts are the roots of the equation $x^2 - hx + k = 0$.

17. (*Construction of the Regular 17-gon*) Draw a circle of radius 2 centered at the origin (0, 0).

(a) Join the point (4, 0) to the point (0, 1) and construct the line ℓ_1 bisecting the angle between this line and the y-axis. Construct the line ℓ_2 perpendicular to ℓ_1 in Figure 2.

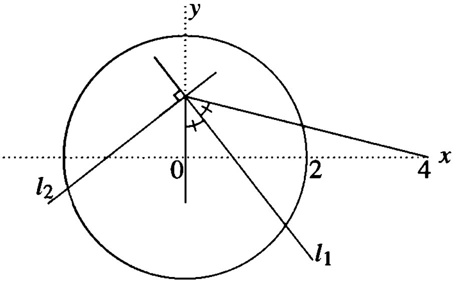


Fig. 2

(b) Using the intersection of ℓ_1 and the x-axis as center and radius equal to the distance to (0, 1), construct the circle C_1 and let $A = (s, 0)$ be the right-hand point of intersection of C_1 with the x-axis. Similarly, let $B = (t, 0)$ denote the right-hand point of intersection of the x-axis and the circle C_2 whose center is the intersection of ℓ_2 and the x-axis and whose radius is equal to the distance to (0, 1) as in Figure 3.

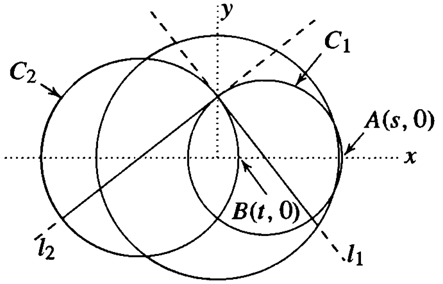


Fig. 3

(c) Construct a perpendicular to the x-axis at the point A and mark off the distance t from (0, 0) to B to construct the point (s, t). Construct the circle with (s, t) and (0, 1) as a diameter and let P denote the right-hand point of intersection of this circle with the x-axis. The perpendicular to the x-axis at P intersects the circle of radius 2 at the second vertex of a regular 17-gon whose first vertex is at (2,0), hence constructs the regular 17-gon by straightedge and compass as in Figure 4.

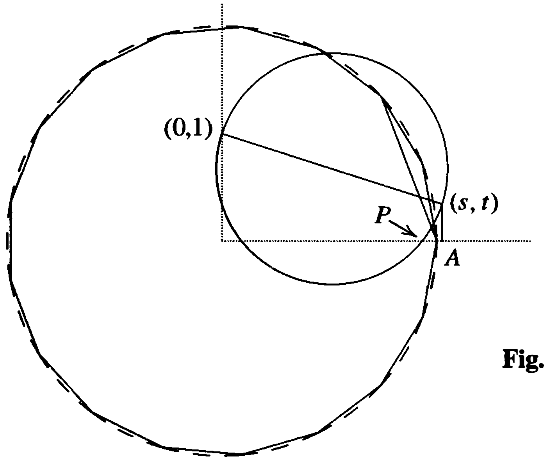


Fig. 4

18. Notation as in the previous exercises.

(a) Prove that ℓ_1 intersects the x-axis in the point $(\eta_1/2, 0)$ and that ℓ_2 intersects the x-axis in the point $(\eta_2/2, 0)$.

(b) Prove that C_1 is the circle having the points $(\eta_1, -1)$ and $(0, 1)$ as diameter. Prove that $s = \eta_1'$. Similarly prove that C_2 is the circle having the points $(\eta_2, -1)$ and $(0, 1)$ as diameter and that $t = \eta_4'$.

(c) Prove that P has coordinates $(\eta_1'', 0)$ and hence that the construction in the previous problem constructs the regular 17-gon by straightedge and compass.

14.6 GALOIS GROUPS OF POLYNOMIALS

Recall that the Galois group of a separable polynomial $f(x) \in F[x]$ is defined to be the Galois group of the splitting field of $f(x)$ over F.

If K is a Galois extension of F then K is the splitting field for some separable polynomial $f(x)$ over F. Any automorphism $\sigma \in \mathrm{Gal}(K/F)$ maps a root of an irreducible factor of $f(x)$ to another root of the irreducible factor and σ is uniquely determined by its action on these roots (since they generate K over F). If we fix a labelling of the roots $\alpha_1, \dots, \alpha_n$ of $f(x)$ we see that any $\sigma \in \mathrm{Gal}(K/F)$ defines a unique permutation of $\alpha_1, \dots, \alpha_n$, hence defines a unique permutation of the subscripts $\{1, 2, \dots, n\}$ (which depends on the fixed labelling of the roots). This gives an injection

$$\mathrm{Gal}(K/F) \hookrightarrow S_n$$

of the Galois group into the symmetric group on n letters which is clearly a homomorphism (both group operations are composition). We may therefore think of Galois groups as subgroups of symmetric groups. Since the degree of the splitting field is the same as the order of the Galois group by the Fundamental Theorem, this explains from the group-theoretic side why the splitting field for a polynomial of degree n over F is of degree at most $n!$ over F (Proposition 13.26).

In general, if the factorization of $f(x)$ into irreducibles is $f(x) = f_1(x) \cdots f_k(x)$ where $f_i(x)$ has degree n_i, $i = 1, 2, \dots, k$, then since the Galois group permutes the roots of the irreducible factors among themselves we have $\mathrm{Gal}(K/F) \le S_{n_1} \times \cdots \times S_{n_k}$.

If $f(x)$ is irreducible, then given any two roots of $f(x)$ there is an automorphism in the Galois group G of $f(x)$ which maps the first root to the second (this follows from our extension Theorem 13.27). Such a group is said to be *transitive* on the roots, i.e., you can get from any given root to any other root by applying some element of G. The fact that the Galois group must be transitive on blocks of roots (namely, the roots of the irreducible factors) can often be helpful in reducing the number of possibilities for the structure of G (cf. the discussion of Galois groups of polynomials of degree 4 below).

Examples

(1) Consider the biquadratic extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$, which is the splitting field of $(x^2-2)(x^2-3)$. Label the roots as $\alpha_1 = \sqrt{2}$, $\alpha_2 = -\sqrt{2}$, $\alpha_3 = \sqrt{3}$ and $\alpha_4 = -\sqrt{3}$. The elements of the Galois group are $\{1, \sigma, \tau, \sigma\tau\}$ where σ maps $\sqrt{2}$ to $-\sqrt{2}$ and fixes $\sqrt{3}$ and τ fixes $\sqrt{2}$ and maps $\sqrt{3}$ to $-\sqrt{3}$. As permutations of the roots for this

labelling we see that σ interchanges the first two and fixes the second two and τ fixes the first two and interchanges the second two, i.e.,

$$\sigma = (12) \qquad \text{and} \qquad \tau = (34)$$

as elements of S_4. Similarly, or by taking the product of these two elements, we see that

$$\sigma\tau = (12)(34) \in S_4.$$

Hence

$$\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}) \cong \{1, (12), (34), (12)(34)\} \subset S_4$$

identifying this Galois group with the Klein-4 subgroup of S_4. Note that if we had changed the labelling of the roots above we would have obtained a different (isomorphic) representation of the Galois group as a subgroup of S_4 (for example, interchanging the second and third roots would have given the subgroup $\{1, (13), (24), (13)(24)\}$).

(2) The Galois group of $x^3 - 2$ acts as permutations on the three roots $\sqrt[3]{2}$, $\rho\sqrt[3]{2}$ and $\rho^2\sqrt[3]{2}$ where ρ is a primitive 3rd root of unity. With this ordering, the generators σ and τ we have defined earlier give the permutations

$$\sigma = (123) \qquad \tau = (23)$$

which gives

$$\{1, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\} = \{1, (123), (132), (23), (13), (12)\} = S_3,$$

in this case the full symmetric group on 3 letters.

Recall that every finite group is isomorphic to a subgroup of some symmetric group S_n. It is an open problem to determine whether every finite group appears as the Galois group for some polynomial over $\mathbb{Q}$. We have seen in the last section that every abelian group is a Galois group over $\mathbb{Q}$ (for some subfield of a cyclotomic field). We shall explicitly determine the Galois groups for polynomials of small degree (≤ 4) below which will in particular show that every subgroup of S_4 arises as a Galois group.

We first introduce some definitions and show that the "general" polynomial of degree n has S_n as Galois group (so the second example above should be viewed as "typical").

Definition. Let $x_1, x_2, \ldots, x_n$ be indeterminates. The *elementary symmetric functions* $s_1, s_2, \ldots, s_n$ are defined by

$$\begin{aligned} s_1 &= x_1 + x_2 + \cdots + x_n \\ s_2 &= x_1x_2 + x_1x_3 + \cdots + x_2x_3 + x_2x_4 + \cdots + x_{n-1}x_n \\ &\ \vdots \\ s_n &= x_1x_2 \cdots x_n \end{aligned}$$

i.e., the i^{th} symmetric function s_i of $x_1, x_2, \ldots, x_n$ is the sum of all products of the x_j's taken i at a time.

Definition. The *general polynomial of degree n* is the polynomial

$$(x - x_1)(x - x_2) \cdots (x - x_n)$$

whose roots are the indeterminates $x_1, x_2, \ldots, x_n$.

It is easy to see by induction that the coefficients of the general polynomial of degree n are given by the elementary symmetric functions in the roots:

$$(x - x_1)(x - x_2)\cdots(x - x_n) = x^n - s_1x^{n-1} + s_2x^{n-2} + \cdots + (-1)^n s_n. \quad (14.13)$$

For any field F, the extension $F(x_1, x_2, \ldots, x_n)$ is then a Galois extension of the field $F(s_1, s_2, \ldots, s_n)$ since it is the splitting field of the general polynomial of degree n .

If $\sigma \in S_n$ is any permutation of $\{1, 2, \ldots, n\}$, then σ acts on the rational functions in $F(x_1, x_2, \ldots, x_n)$ by permuting the subscripts of the variables $x_1, x_2, \ldots, x_n$. It is clear that this gives an automorphism of $F(x_1, x_2, \ldots, x_n)$. Identifying $\sigma \in S_n$ with this automorphism of $F(x_1, x_2, \ldots, x_n)$ identifies S_n as a subgroup of $\mathrm{Aut}(F(x_1, x_2, \ldots, x_n))$.

The elementary symmetric functions $s_1, s_2, \ldots, s_n$ are fixed under any permutation of their subscripts (this is the reason they are called *symmetric*), which shows that the subfield $F(s_1, s_2, \ldots, s_n)$ is contained in the fixed field of S_n. By the Fundamental Theorem of Galois Theory, the fixed field of S_n has index precisely $n!$ in $F(x_1, x_2, \ldots, x_n)$. Since $F(x_1, x_2, \ldots, x_n)$ is the splitting field over $F(s_1, s_2, \ldots, s_n)$ of the polynomial of degree n in (13), we have

$$[F(x_1, x_2, \ldots, x_n) : F(s_1, s_2, \ldots, s_n)] \leq n!\,. \quad (14.14)$$

It follows that we actually have equality and that $F(s_1, s_2, \ldots, s_n)$ is precisely the fixed field of S_n. This proves the following result.

Proposition 30. The fixed field of the symmetric group S_n acting on the field of rational functions in n variables $F(x_1, x_2, \ldots, x_n)$ is the field of rational functions in the elementary symmetric functions $F(s_1, s_2, \ldots, s_n)$.

Definition. A rational function $f(x_1, x_2, \ldots, x_n)$ is called *symmetric* if it is not changed by any permutation of the variables $x_1, x_2, \ldots, x_n$.

Corollary 31. *(Fundamental Theorem on Symmetric Functions)* Any symmetric function in the variables $x_1, x_2, \ldots, x_n$ is a rational function in the elementary symmetric functions $s_1, s_2, \ldots, s_n$.

Proof: A symmetric function lies in the fixed field of S_n above, hence is a rational function in $s_1, \ldots, s_n$.

This corollary explains why these are called the *elementary* symmetric functions.

Remark: If $f(x_1, \ldots, x_n)$ is a *polynomial* in $x_1, x_2, \ldots, x_n$ which is symmetric then it can be seen that f is actually a polynomial in $s_1, s_2, \ldots, s_n$, which strengthens the statement of the corollary. It is in fact true that a symmetric polynomial whose coefficients lie in R, where R is any commutative ring with identity, is a polynomial in the elementary symmetric functions with coefficients in R. A proof of this fact is implicit in the algorithm outlined in the exercises for writing a symmetric polynomial as a polynomial in the elementary symmetric functions.

Examples

(1) The expression $(x_1 - x_2)^2$ is symmetric in x_1, x_2. We have

$$(x_1 - x_2)^2 = (x_1 + x_2)^2 - 4x_1x_2 = s_1^2 - 4s_2,$$

a polynomial in the elementary symmetric functions.

(2) The polynomial $x_1^2 + x_2^2 + x_3^2$ is symmetric in x_1, x_2, x_3, and in this case we have

$$\begin{aligned} x_1^2 + x_2^2 + x_3^2 &= (x_1 + x_2 + x_3)^2 - 2(x_1x_2 + x_1x_3 + x_2x_3) \\ &= s_1^2 - 2s_2. \end{aligned}$$

(3) The polynomial $x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2$ is symmetric. Since

$$\begin{aligned} (x_1x_2 + x_1x_3 + x_2x_3)^2 &= x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2 + 2(x_1^2x_2x_3 + x_2^2x_1x_3 + x_3^2x_1x_2) \\ &= x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2 + 2x_1x_2x_3(x_1 + x_2 + x_3) \end{aligned}$$

we have

$$x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2 = s_2^2 - 2s_1s_3.$$

Suppose now we *start* with the general polynomial

$$x^n - s_1x^{n-1} + s_2x^{n-2} + \cdots + (-1)^n s_n$$

over the field $F(s_1, s_2, \dots, s_n)$ where we view the $s_i, i = 1, 2, \dots, n$ as indeterminates. If we define the roots of this polynomial to be $x_1, x_2, \dots, x_n$ then the s_i are precisely the elementary symmetric functions in the roots $x_1, \dots, x_n$. Moreover, these roots are indeterminates as well in the sense that there are no polynomial relations over F between them. For suppose $p(t_1, \dots, t_n)$ is a nonzero polynomial in n variables with coefficients in F such that $p(x_1, \dots, x_n) = 0$. Then the product, $\widetilde{p}$, over all σ in S_n of $p(t_{\sigma(1)}, \dots, t_{\sigma(n)})$ is a nonzero symmetric polynomial with $\widetilde{p}(x_1, \dots, x_n) = 0$. This gives a nonzero polynomial relation over F among $s_1, \dots, s_n$, a contradiction. Conversely, if the roots of a polynomial $f(x)$ are independent indeterminates over F, then so are the coefficients of $f(x)$ — cf. the beginning of Section 9. Thus defining the general polynomial over F as having indeterminate roots or indeterminate coefficients is equivalent. From this point of view our result can be stated in the following form.

Theorem 32. The general polynomial

$$x^n - s_1x^{n-1} + s_2x^{n-2} + \cdots + (-1)^n s_n$$

over the field $F(s_1, s_2, \dots, s_n)$ is separable with Galois group S_n.

This result says that if there are no relations among the coefficients of a polynomial of degree n (which is what we mean when we say the s_i are indeterminates above) then the Galois group of this polynomial over the field generated by its coefficients is the full symmetric group S_n. Loosely speaking, this means that the "generic" polynomial of degree n will have S_n as Galois group. Note, however, that over finite fields every polynomial has a *cyclic* Galois group (all extensions of finite fields are cyclic), so that "generic" polynomials in this sense do not exist. Over $\mathbb{Q}$ one can make precise the notion of "generic" polynomial and then it is true that most polynomials have the full symmetric group as Galois group.

For $n \geq 5$ there is only one normal subgroup of S_n, namely the subgroup A_n of index 2. Hence in general there is only one normal subfield of $F(x_1, x_2, \ldots, x_n)$ containing $F(s_1, s_2, \ldots, s_n)$ and it is an extension of degree 2.

Definition. Define the *discriminant* D of $x_1, x_2, \ldots, x_n$ by the formula

$$D = \prod_{i<j}(x_i - x_j)^2.$$

Define the discriminant of a polynomial to be the discriminant of the roots of the polynomial.

The discriminant D is a symmetric function in $x_1, \ldots, x_n$, hence is an element of $K = F(s_1, s_2, \ldots, s_n)$.

When we first defined the alternating group A_n we saw that a permutation $\sigma \in S_n$ is an element of the subgroup A_n if and only if σ fixes the product

$$\sqrt{D} = \prod_{i<j}(x_i - x_j) \in \mathbb{Z}[x_1, x_2, \ldots, x_n].$$

It follows (by the Fundamental Theorem) that if F has characteristic different from 2 then $\sqrt{D}$ generates the fixed field of A_n and generates a quadratic extension of K. This proves the following proposition.

Proposition 33. If $\text{ch}(F) \neq 2$ then the permutation $\sigma \in S_n$ is an element of A_n if and only if it fixes the square root of the discriminant D.

We now consider the Galois groups of separable polynomials of small degree (≤ 4) over a field F which we assume is of characteristic different from 2 and 3. Note that over $\mathbb{Q}$ or over a finite field (or, more generally, over any perfect field) the splitting field of an arbitrary polynomial $f(x)$ is the same as the splitting field for the product of the irreducible factors of $f(x)$ taken precisely once, which is a separable polynomial.

If the roots of the polynomial $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ are $\alpha_1, \alpha_2, \ldots, \alpha_n$, then the discriminant of $f(x)$ is[2]

$$D = \prod_{i<j}(\alpha_i - \alpha_j)^2.$$

Note that $D = 0$ if and only if $f(x)$ is not separable, i.e., if the roots $\alpha_1, \ldots, \alpha_n$ are not distinct. Recall that over a perfect field (e.g., $\mathbb{Q}$ or a finite field) this implies $f(x)$ is reducible since every irreducible polynomial over a perfect field is separable.

The discriminant D is symmetric in the roots of $f(x)$, hence is fixed by all the automorphisms of the Galois group of $f(x)$. By the Fundamental Theorem it follows that

[2] If $f(x) = a_nx^n + \cdots + a_0$ is not monic then its discriminant is defined to be a_n^{2n-2} times the D defined above.

$D \in F$. The discriminant can in general be written as a polynomial in the coefficients of $f(x)$ (by Corollary 31) which are fairly complicated for larger degrees (we shall give formulas for $n \leq 4$ below). Finally, note that since

$$\sqrt{D} = \prod_{i<j}(\alpha_i - \alpha_j)$$

we have the useful fact that $\sqrt{D}$ is always contained in the splitting field for $f(x)$.

If the roots of $f(x)$ are distinct, fix some ordering of the roots and view the Galois group of $f(x)$ as a subgroup of S_n as above.

Proposition 34. The Galois group of $f(x) \in F[x]$ is a subgroup of A_n if and only if the discriminant $D \in F$ is the square of an element of F.

Proof: This is a restatement of Proposition 33 in this case. The Galois group is contained in A_n if and only if every element of the Galois group fixes

$$\sqrt{D} = \prod_{i<j}(\alpha_i - \alpha_j)$$

i.e., if and only if $\sqrt{D} \in F$.

This property, together with the fact that $D = 0$ determines the presence of multiple roots, is the reason D is called the *discriminant*.

Polynomials of Degree 2

Consider the polynomial $x^2 + ax + b$ with roots α, β. The discriminant D for this polynomial is $(\alpha - \beta)^2$, which can be written as a polynomial in the elementary symmetric functions of the roots. We did this in Example 1 above:

$$D = s_1^2 - 4s_2 = (-a)^2 - 4(b) = a^2 - 4b,$$

the usual discriminant for this quadratic.

The polynomial is separable if and only if $a^2 - 4b \neq 0$. The Galois group is a subgroup of S_2, the cyclic group of order 2 and is trivial (i.e., A_2 in this case) if and only if $a^2 - 4b$ is a rational square, which completely determines the possible Galois groups.

Note that this restates results we obtained previously by explicitly solving for the roots: if the polynomial is reducible (namely D is a square in F), then the Galois group is trivial (the splitting field is just F), while if the polynomial is irreducible the Galois group is isomorphic to $\mathbb{Z}/2\mathbb{Z}$ since the splitting field is the quadratic extension $F(\sqrt{D})$.

Polynomials of degree 3

Suppose the cubic polynomial is

$$f(x) = x^3 + ax^2 + bx + c. \tag{14.15}$$

If we make the substitution $x = y - a/3$ the polynomial becomes

$$g(y) = y^3 + py + q \tag{14.16}$$

where

$$p = \frac{1}{3}(3b - a^2) \qquad q = \frac{1}{27}(2a^3 - 9ab + 27c). \tag{14.17}$$

The splitting fields for these two polynomials are the same since their roots differ by the constant $a/3 \in F$ and since the formula for the discriminant involves the *differences* of roots, we see that these two polynomials also have the *same* discriminant.

Let the roots of the polynomial in (16) be α, β, and γ. We first compute the discriminant of this polynomial in terms of p and q. Note that

$$g(y) = (y - \alpha)(y - \beta)(y - \gamma)$$

so that if we differentiate we have

$$D_y g(y) = (y - \alpha)(y - \beta) + (y - \alpha)(y - \gamma) + (y - \beta)(y - \gamma).$$

Then

$$\begin{aligned} D_y g(\alpha) &= (\alpha - \beta)(\alpha - \gamma) \\ D_y g(\beta) &= (\beta - \alpha)(\beta - \gamma) \\ D_y g(\gamma) &= (\gamma - \alpha)(\gamma - \beta). \end{aligned}$$

Taking the product we see that

$$D = [(\alpha - \beta)(\alpha - \gamma)(\beta - \gamma)]^2 = -D_y g(\alpha) D_y g(\beta) D_y g(\gamma).$$

Since $D_y g(y) = 3y^2 + p$, we have

$$\begin{aligned} -D &= (3\alpha^2 + p)(3\beta^2 + p)(3\gamma^2 + p) \\ &= 27\alpha^2\beta^2\gamma^2 + 9p(\alpha^2\beta^2 + \alpha^2\gamma^2 + \beta^2\gamma^2) + 3p^2(\alpha^2 + \beta^2 + \gamma^2) + p^3. \end{aligned}$$

The corresponding expressions in the elementary symmetric functions of the roots were determined in Examples 2 and 3 above. Note that here $s_1 = 0$, $s_2 = p$ and $s_3 = -q$. We obtain

$$-D = 27(-q)^2 + 9p(p^2) + 3p^2(-2p) + p^3$$

so that

$$D = -4p^3 - 27q^2. \tag{14.18}$$

This is the same as the discriminant of $f(x)$ in (15). Expressing D in terms of a, b, c using (17) we obtain

$$D = a^2b^2 - 4b^3 - 4a^3c - 27c^2 + 18abc \tag{14.18$'$}$$

(Galois Group of a Cubic)

a. If the cubic polynomial $f(x)$ is reducible, then it splits either into three linear factors or into a linear factor and an irreducible quadratic. In the first case the Galois group is trivial and in the second case the Galois group is of order 2.

b. If the cubic polynomial $f(x)$ is irreducible then a root of $f(x)$ generates an extension of degree 3 over F, so the degree of the splitting field over F is divisible by 3. Since the Galois group is a subgroup of S_3, there are only two possibilities, namely

A_3 or S_3. The Galois group is A_3 (i.e., cyclic of order 3) if and only if the discriminant D in (18) is a square.

Explicitly, if D is the square of an element of F, then the splitting field of the irreducible cubic $f(x)$ is obtained by adjoining any single root of $f(x)$ to F. The resulting field is Galois over F of degree 3 with a cyclic group of order 3 as Galois group. If D is not the square of an element of F then the splitting field of $f(x)$ is of degree 6 over F, hence is the field $F(\theta, \sqrt{D})$ for any one of the roots θ of $f(x)$. This extension is Galois over F with Galois group S_3 (generators are given by σ, which takes θ to one of the other roots of $f(x)$ and fixes $\sqrt{D}$, and τ, which takes $\sqrt{D}$ to $-\sqrt{D}$ and fixes θ).

We see that in both cases the splitting field for the irreducible cubic $f(x)$ is obtained by adjoining $\sqrt{D}$ and a root of $f(x)$ to F.

We shall give explicit formulas for the roots of (16) (*Cardano's Formulas*) in the next section after introducing the notion of a *Lagrange Resolvent*.

Polynomials of Degree 4

Let the quartic polynomial be

$$f(x) = x^4 + ax^3 + bx^2 + cx + d$$

which under the substitution $x = y - a/4$ becomes the quartic

$$g(y) = y^4 + py^2 + qy + r$$

with

$$\begin{aligned} p &= \frac{1}{8}(-3a^2 + 8b) \\ q &= \frac{1}{8}(a^3 - 4ab + 8c) \\ r &= \frac{1}{256}(-3a^4 + 16a^2b - 64ac + 256d). \end{aligned}$$

Let the roots of $g(y)$ be α_1, α_2, α_3, and α_4 and let G denote the Galois group for the splitting field of $g(y)$ (or of $f(x)$).

Suppose first that $g(y)$ is reducible. If $g(y)$ splits into a linear and a cubic, then G is the Galois group of the cubic, which we determined above. Suppose then that $g(y)$ splits into two irreducible quadratics. Then the splitting field is the extension $F(\sqrt{D_1}, \sqrt{D_2})$ where D_1 and D_2 are the discriminants of the two quadratics. If D_1 and D_2 do not differ by a square factor then this extension is a biquadratic extension and G is isomorphic to the Klein 4-subgroup of S_4. If D_1 is a square times D_2 then this extension is a quadratic extension and G is isomorphic to $\mathbb{Z}/2\mathbb{Z}$.

We are reduced to the situation where $g(y)$ is irreducible. In this case recall that the Galois group is transitive on the roots, i.e., it is possible to get from a given root to any other root by applying some automorphism of the Galois group. Examining the possibilities we see that the only transitive subgroups of S_4, hence the only possibilities for our Galois group G, are the groups

S_4, A_4

$D_8 = \{1, (1324), (12)(34), (1423), (13)(24), (14)(23), (12), (34)\}$ and its conjugates

$V = \{1, (12)(34), (13)(24), (14)(23)\}$

$C = \{1, (1234), (13)(24), (1432)\}$ and its conjugates.

(D_8 is the dihedral group, a Sylow 2-subgroup of S_4, with 3 (isomorphic) conjugate subgroups in S_4, V is the Klein 4-subgroup of S_4, normal in S_4, and C is a cyclic group, with 3 (isomorphic) conjugates in S_4).

Consider the elements

$$\begin{aligned} \theta_1 &= (\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4) \\ \theta_2 &= (\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4) \\ \theta_3 &= (\alpha_1 + \alpha_4)(\alpha_2 + \alpha_3) \end{aligned}$$

in the splitting field for $g(y)$. These elements are permuted amongst themselves by the permutations in S_4. The stabilizer of θ_1 in S_4 is the dihedral group D_8. The stabilizers in S_4 of θ_2 and θ_3 are the conjugate dihedral subgroups of order 8. The subgroup of S_4 which stabilizes all three of these elements is the intersection of these subgroups, namely the Klein 4-group V.

Since S_4 merely permutes $\theta_1, \theta_2, \theta_3$ it follows that the elementary symmetric functions in the θ's are fixed by all the elements of S_4, hence are in F. An elementary computation in symmetric functions shows that these elementary symmetric functions are $2p$, $p^2 - 4r$, and $-q^2$, which shows that $\theta_1, \theta_2, \theta_3$ are the roots of

$$h(x) = x^3 - 2px^2 + (p^2 - 4r)x + q^2$$

called the *resolvent cubic* for the quartic $g(y)$. Since

$$\begin{aligned} \theta_1 - \theta_2 &= \alpha_1\alpha_3 + \alpha_2\alpha_4 - \alpha_1\alpha_2 - \alpha_3\alpha_4 \\ &= -(\alpha_1 - \alpha_4)(\alpha_2 - \alpha_3) \end{aligned}$$

and similarly

$$\begin{aligned} \theta_1 - \theta_3 &= -(\alpha_1 - \alpha_3)(\alpha_2 - \alpha_4) \\ \theta_2 - \theta_3 &= -(\alpha_1 - \alpha_2)(\alpha_3 - \alpha_4) \end{aligned}$$

we see that the discriminant of the resolvent cubic is the *same* as the discriminant of the quartic $g(y)$, hence also as the discriminant of the quartic $f(x)$. Using our formula for the discriminant of the cubic, we can easily compute the discriminant in terms of p, q, r:

$$D = 16p^4r - 4p^3q^2 - 128p^2r^2 + 144pq^2r - 27q^4 + 256r^3$$

from which one can give the formula for D in terms of a, b, c, d:

$$\begin{aligned} D = &- 128b^2d^2 - 4a^3c^3 + 16b^4d - 4b^3c^2 - 27a^4d^2 + 18abc^3 \\ &+ 144a^2bd^2 - 192acd^2 + a^2b^2c^2 - 4a^2b^3d - 6a^2c^2d \\ &+ 144bc^2d + 256d^3 - 27c^4 - 80ab^2cd + 18a^3bcd. \end{aligned}$$

The splitting field for the resolvent cubic is a subfield of the splitting field of the quartic, so the Galois group of the resolvent cubic is a quotient of G. Hence knowing the action of the Galois group on the roots of the resolvent cubic $h(x)$ gives information about the Galois group of $g(y)$, as follows:

(Galois group of a quartic)

a. Suppose first that the resolvent cubic is irreducible. If D is not a square, then G is not contained in A_4 and the Galois group of the resolvent cubic is S_3, which implies that the degree of the splitting field for $g(y)$ is divisible by 6. The only possibility is then $G = S_4$.

b. If the resolvent cubic is irreducible and D is a square, then G is a subgroup of A_4 and 3 divides the order of G (the Galois group of the resolvent cubic is A_3). The only possibility is $G = A_4$.

c1. We are left with the case where the resolvent cubic is reducible. The first possibility is that $h(x)$ has 3 roots in F (i.e., splits completely). Since each of the elements $\theta_1, \theta_2, \theta_3$ is in F, every element of G fixes all three of these elements, which means $G \subseteq V$. The only possibility is $G = V$.

c2. If $h(x)$ splits into a linear and a quadratic, then precisely one of $\theta_1, \theta_2, \theta_3$ is in F, say θ_1. Then G stabilizes θ_1 but not θ_2 and θ_3, so we have $G \subseteq D_8$ and $G \not\subseteq V$. This leaves two possibilities: $G = D_8$ or $G = C$. One way to distinguish between these is to observe that $F(\sqrt{D})$ is the fixed field of the elements of G in A_4. For the two cases being considered, we have $D_8 \cap A_4 = V$, $C \cap A_4 = \{1, (13)(24)\}$. The first group is transitive on the roots of $g(y)$, the second is not. It follows that the first case occurs if and only if $g(y)$ is irreducible over $F(\sqrt{D})$. We may therefore determine G completely by factoring $g(y)$ in $F(\sqrt{D})$, and so completely determine the Galois group in all cases. (cf. the exercises following and in the next section, where it is shown that over $\mathbb{Q}$ the Galois group cannot be cyclic of degree 4 if D is not the sum of two squares — so in particular if $D < 0$.)

We shall give explicit formulas for the roots of a quartic polynomial at the end of the next section.

The Fundamental Theorem of Algebra

We end this section with two proofs of the Fundamental Theorem of Algebra. We need two facts regarding the field $\mathbb{C}$:

(a) Every polynomial with real coefficients of odd degree has a root in the reals. Equivalently, there are no nontrivial finite extensions of $\mathbb{R}$ of odd degree.

(b) Quadratic polynomials with coefficients in $\mathbb{C}$ have roots in $\mathbb{C}$. Equivalently, there are no quadratic extensions of $\mathbb{C}$.

The first result follows from the Intermediate Value Theorem in calculus, since the graph of a monic polynomial $f(x) \in \mathbb{R}[x]$ of odd degree is negative for large negative values of x and positive for large positive values of x, hence crosses the axis somewhere. The equivalence with the second statement follows since a finite extension of $\mathbb{R}$ is a simple extension and the minimal polynomial of a primitive element would have odd degree, hence would be both irreducible over $\mathbb{R}$ and have a root in $\mathbb{R}$, hence must be of degree 1.

The second result follows by a direct computation. By the quadratic formula it suffices to show that every complex number $\alpha = a + bi$, $a, b \in \mathbb{R}$, has a square root in $\mathbb{C}$. Write $\alpha = re^{i\theta}$ for some $r \geq 0$ and some $\theta \in [0, 2\pi)$. Then $\sqrt{r}e^{i\theta/2}$ is a square root of α. (Explicitly, let $c \in \mathbb{R}$ be a square root of the real number $\dfrac{a + \sqrt{a^2 + b^2}}{2}$ and let $d \in \mathbb{R}$ be a square root of the real number $\dfrac{-a + \sqrt{a^2 + b^2}}{2}$ where the signs of the two square roots are chosen so that cd has the same sign as b. Then multiplying out we see that $(c + di)^2 = a + bi$.)

Theorem 35. *(Fundamental Theorem of Algebra)* Every polynomial $f(x) \in \mathbb{C}[x]$ of degree n has precisely n roots in $\mathbb{C}$ (counted with multiplicity). Equivalently, $\mathbb{C}$ is algebraically closed.

Proof: I. It suffices to prove that every polynomial $f(x) \in \mathbb{C}[x]$ has a root in $\mathbb{C}$. Let τ denote the automorphism complex conjugation. If $f(x)$ has no root in $\mathbb{C}$ then neither does the conjugate polynomial $\bar{f}(x) = \tau f(x)$ obtained by applying τ to the coefficients of $f(x)$ (since its roots are the conjugates of the roots of $f(x)$). The product $f(x)\bar{f}(x)$ has coefficients which are invariant under complex conjugation, hence has real coefficients. It suffices then to prove that a polynomial with real coefficients has a root in $\mathbb{C}$.

Suppose that $f(x)$ is a polynomial of degree n with real coefficients and write $n = 2^k m$ where m is odd. We prove that $f(x)$ has a root in $\mathbb{C}$ by induction on k. For $k = 0$, $f(x)$ has odd degree and by (a) above $f(x)$ has a root in $\mathbb{R}$ so we are done. Suppose now that $k \geq 1$. Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be the roots of $f(x)$ and set $K = \mathbb{R}(\alpha_1, \alpha_2, \ldots, \alpha_n, i)$. Then K is a Galois extension of $\mathbb{R}$ containing $\mathbb{C}$ and the roots of $f(x)$. For any $t \in \mathbb{R}$ consider the polynomial

$$L_t = \prod_{1 \leq i < j \leq n} [x - (\alpha_i + \alpha_j + t\alpha_i\alpha_j)].$$

Any automorphism of $K/\mathbb{R}$ permutes the terms in this product so the coefficients of L_t are invariant under all the elements of $\mathrm{Gal}(K/\mathbb{R})$. Hence L_t is a polynomial with real coefficients. The degree of L_t is

$$\frac{n(n-1)}{2} = 2^{k-1}m(2^k m - 1) = 2^{k-1}m'$$

where m' is odd (since $k \geq 1$). The power of 2 in this degree is therefore less than k, so by induction the polynomial L_t has a root in $\mathbb{C}$. Hence for each $t \in \mathbb{R}$ one of the elements $\alpha_i + \alpha_j + t\alpha_i\alpha_j$ for some i, j $(1 \leq i < j \leq n)$ is an element of $\mathbb{C}$. Since there are infinitely many choices for t and only finitely many values of i and j we see that for some i and j (say, $i = 1$ and $j = 2$) there are distinct real numbers s and t with

$$\alpha_1 + \alpha_2 + s\alpha_1\alpha_2 \in \mathbb{C} \qquad \alpha_1 + \alpha_2 + t\alpha_1\alpha_2 \in \mathbb{C}.$$

Since $s \neq t$ it follows that $a = \alpha_1 + \alpha_2 \in \mathbb{C}$ and $b = \alpha_1\alpha_2 \in \mathbb{C}$. But then α_1 and α_2 are the roots of the quadratic $x^2 - ax + b$ with coefficients in $\mathbb{C}$, hence are elements of $\mathbb{C}$ by (b) above, completing the proof.

II. The second proof again uses (a) and (b) above, but replaces the computations with the polynomials L_t above with a simple group-theoretic argument involving the nilpotency of a Sylow 2-subgroup of the Galois group:

Let $f(x)$ be a polynomial of degree n with real coefficients and let K be the splitting field of $f(x)$ over $\mathbb{R}$. Then $K(i)$ is a Galois extension of $\mathbb{R}$. Let G denote its Galois group and let P_2 denote a Sylow 2-subgroup of G. The fixed field of P_2 is an extension of $\mathbb{R}$ of odd degree, hence by (a) is trivial.

It follows that $\text{Gal}(K(i)/\mathbb{C})$ is a 2-group. Since 2-groups have subgroups of all orders (recall this is true of a finite p-group for any prime p, cf. Theorem 6.1), if this group is nontrivial, there would exist a quadratic extension of $\mathbb{C}$, impossible by (b), completing the proof.

The Fundamental Theorem of Algebra was first rigorously proved by Gauss in 1816 (his doctoral dissertation in 1798 provides a proof using geometric considerations requiring some topological justification). The first proof above is essentially due to Laplace in 1795 (hence the reason for naming the polynomials L_t). The reason Laplace's proof was deemed unacceptable was that he assumed the existence of a splitting field for polynomials (i.e., that the roots existed *somewhere* in *some* field), which had not been established at that time. The elegant second proof is a simplification due to Artin.

EXERCISES

1. Show that a cubic with a multiple root has a linear factor. Is the same true for quartics?
2. Determine the Galois groups of the following polynomials:
   (a) $x^3 - x^2 - 4$
   (b) $x^3 - 2x + 4$
   (c) $x^3 - x + 1$
   (d) $x^3 + x^2 - 2x - 1$.
3. Prove for any $a, b \in \mathbb{F}_{p^n}$ that if $x^3 + ax + b$ is irreducible then $-4a^3 - 27b^2$ is a square in $\mathbb{F}_{p^n}$.
4. Determine the Galois group of $x^4 - 25$.
5. Determine the Galois group of $x^4 + 4$.
6. Determine the Galois group of $x^4 + 3x^3 - 3x - 2$.
7. Determine the Galois group of $x^4 + 2x^2 + x + 3$.
8. Determine the Galois group of $x^4 + 8x + 12$.
9. Determine the Galois group of $x^4 + 4x - 1$ (cf. Exercise 19).
10. Determine the Galois group of $x^5 + x - 1$.
11. Let F be an extension of $\mathbb{Q}$ of degree 4 that is not Galois over $\mathbb{Q}$. Prove that the Galois closure of F has Galois group either S_4, A_4 or the dihedral group D_8 of order 8. Prove that the Galois group is dihedral if and only if F contains a quadratic extension of $\mathbb{Q}$.
12. Prove that an extension F of $\mathbb{Q}$ of degree 4 can be generated by the root of an irreducible biquadratic $x^4 + ax^2 + b$ over $\mathbb{Q}$ if and only if F contains a quadratic extension of $\mathbb{Q}$.

13. **(a)** Let $\pm\alpha, \pm\beta$ denote the roots of the polynomial $f(x) = x^4 + ax^2 + b \in \mathbb{Z}[x]$. Prove that $f(x)$ is irreducible if and only if $\alpha^2, \alpha \pm \beta$ are not elements of $\mathbb{Q}$.[3]

(b) Suppose $f(x)$ is irreducible and let G be the Galois group of $f(x)$. Prove that

(i) $G \cong V$, the Klein 4-group, if and only if b is a square in $\mathbb{Q}$ if and only if $\alpha\beta \in \mathbb{Q}$ is rational.

(ii) $G \cong C$, the cyclic group of order 4, if and only if $b(a^2 - 4b)$ is a square in $\mathbb{Q}$ if and only if $\mathbb{Q}(\alpha\beta) = \mathbb{Q}(\alpha^2)$.

(iii) $G \cong D_8$, the dihedral group of order 8, if and only if b and $b(a^2 - 4b)$ are not squares in $\mathbb{Q}$ if and only if $\alpha\beta \notin \mathbb{Q}(\alpha^2)$.

14. Prove the polynomial $x^4 - px^2 + q \in \mathbb{Q}[x]$ is irreducible for any distinct odd primes p and q and has as Galois group the dihedral group of order 8.[4]

15. Prove the polynomial $x^4 + px + p \in \mathbb{Q}[x]$ is irreducible for every prime p and for $p \neq 3, 5$ has Galois group S_4. Prove the Galois group for $p = 3$ is dihedral of order 8 and for $p = 5$ is cyclic of order 4.[5]

16. Determine the Galois group over $\mathbb{Q}$ of the polynomial $x^4 + 8x^2 + 8x + 4$. Determine which of the subfields of this field are Galois over $\mathbb{Q}$ and for those which are Galois determine a polynomial $f(x) \in \mathbb{Q}[x]$ for which they are the splitting field over $\mathbb{Q}$.

17. Find the Galois group of $x^4 - 7$ over $\mathbb{Q}$ explicitly as a permutation group on the roots.

18. Let θ be a root of $x^3 - 3x + 1$. Prove that the splitting field of this polynomial is $\mathbb{Q}(\theta)$ and that the Galois group is cyclic of order 3. In particular the other roots of this polynomial can be written in the form $a + b\theta + c\theta^2$ for some $a, b, c \in \mathbb{Q}$. Determine the other roots explicitly in terms of θ.

19. Let $f(x)$ be an irreducible polynomial of degree 4 in $\mathbb{Q}[x]$ with discriminant D. Let K denote the splitting field of $f(x)$, viewed as a subfield of the complex numbers $\mathbb{C}$.

(a) Prove that $\mathbb{Q}(\sqrt{D}) \subset K$.

(b) Let τ denote complex conjugation and let τ_K denote the restriction of complex conjugation to K. Prove that τ_K is an element of $\text{Gal}(K/\mathbb{Q})$ of order 1 or 2 depending on whether every element of K is real or not.

(c) Prove that if $D < 0$ then K cannot be cyclic of degree 4 over $\mathbb{Q}$ (i.e., $\text{Gal}(K/\mathbb{Q})$ cannot be a cyclic group of order 4).

(d) Prove generally that $\mathbb{Q}(\sqrt{D})$ for squarefree $D < 0$ is not a subfield of a cyclic quartic field (cf. also Exercise 19 of Section 7).

20. Determine the Galois group of $(x^3 - 2)(x^3 - 3)$ over $\mathbb{Q}$. Determine all the subfields which contain $\mathbb{Q}(\rho)$ where ρ is a primitive 3rd root of unity.

21. Let $G \leq S_n$ be a subgroup of the symmetric group and suppose $\sigma_1, \ldots, \sigma_k$ are generators for G. If the function $f(x_1, x_2, \ldots, x_n)$ is fixed by the generators σ_i show it is fixed by G.

22. (*Newton's Formulas*) Let $f(x)$ be a monic polynomial of degree n with roots $\alpha_1, \ldots, \alpha_n$. Let s_i be the elementary symmetric function of degree i in the roots and define $s_i = 0$ for $i > n$. Let $p_i = \alpha_1^i + \cdots + \alpha_n^i$, $i \geq 0$, be the sum of the i^{th} powers of the roots of $f(x)$.

[3] cf. the note *An Elementary Test for the Galois Group of a Quartic Polynomial*, Luise-Charlotte Kappe and Bette Warren, Amer. Math. Monthly, 96(1989), pp. 133–137.

[4] Ibid.

[5] Ibid.

Prove *Newton's Formulas*:

$$\begin{aligned}
&p_1 - s_1 = 0\\
&p_2 - s_1 p_1 + 2s_2 = 0\\
&p_3 - s_1 p_2 + s_2 p_1 - 3s_3 = 0\\
&\quad\vdots\\
&p_i - s_1 p_{i-1} + s_2 p_{i-2} - \cdots + (-1)^{i-1} s_{i-1} p_1 + (-1)^i i s_i = 0
\end{aligned}$$

23. **(a)** If $x + y + z = 1$, $x^2 + y^2 + z^2 = 2$ and $x^3 + y^3 + z^3 = 3$, determine $x^4 + y^4 + z^4$.

(b) Prove generally that x, y, z are not rational but that $x^n + y^n + z^n$ is rational for every positive integer n.

24. Prove that an $n \times n$ matrix A over a field of characteristic 0 is nilpotent if and only if the trace of A^k is 0 for all $k \geq 0$.

25. Prove that two $n \times n$ matrices A and B over a field of characteristic 0 have the same characteristic polynomial if and only if the trace of A^k equals the trace of B^k for all $k \geq 0$.

26. Use the fact that the trace of AB is the same as the trace of BA for any two $n \times n$ matrices A and B to show that AB and BA have the same characteristic polynomial over a field of characteristic 0 (the same result is true over a field of arbitrary characteristic).

27. Let $f(x)$ be a monic polynomial of degree n with roots $\alpha_1, \alpha_2, \ldots, \alpha_n$.

(a) Show that the discriminant D of $f(x)$ is the square of the Vandermonde determinant

$$\begin{vmatrix} 1 & \alpha_1 & \alpha_1^2 & \cdots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \alpha_2^2 & \cdots & \alpha_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_n & \alpha_n^2 & \cdots & \alpha_n^{n-1} \end{vmatrix} = \prod_{i>j} (\alpha_i - \alpha_j).$$

(b) Taking the Vandermonde matrix above, multiplying on the left by its transpose and taking the determinant show that one obtains

$$D = \begin{vmatrix} p_0 & p_1 & p_2 & \cdots & p_{n-1} \\ p_1 & p_2 & p_3 & \cdots & p_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{n-1} & p_n & p_{n+1} & \cdots & p_{2n-2} \end{vmatrix}$$

where $p_i = \alpha_1^i + \cdots + \alpha_n^i$ is the sum of the i^{th} powers of the roots of $f(x)$, which can be computed in terms of the coefficients of $f(x)$ using Newton's formulas above. This gives an efficient procedure for calculating the discriminant of a polynomial.

28. Let α be a root of the irreducible polynomial $f(x) \in F[x]$ and let $K = F(\alpha)$. Let D be the discriminant of $f(x)$. Prove that $D = (-1)^{n(n-1)/2} \mathrm{N}_{K/F}(f'(\alpha))$, where $f'(x) = D_x f(x)$ is the derivative of $f(x)$.

The following exercises describe the *resultant* of two polynomials and in particular provide another efficient method for calculating the discriminant of a polynomial.

29. Let F be a field and let $f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ and $g(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0$ be two polynomials in $F[x]$.

(a) Prove that a necessary and sufficient condition for $f(x)$ and $g(x)$ to have a common root (or, equivalently, a common divisor in $F[x]$) is the existence of a polynomial

$a(x) \in F[x]$ of degree at most $m-1$ and a polynomial $b(x) \in F[x]$ of degree at most $n-1$ with $a(x)f(x) = b(x)g(x)$.

(b) Writing $a(x)$ and $b(x)$ explicitly as polynomials show that equating coefficients in the equation $a(x)f(x) = b(x)g(x)$ gives a system of $n+m$ linear equations for the coefficients of $a(x)$ and $b(x)$. Prove that this system has a nontrivial solution (hence $f(x)$ and $g(x)$ have a common zero) if and only if the determinant

$$R(f,g) = \begin{vmatrix} a_n & a_{n-1} & \cdots & a_0 & & & & \\ & a_n & a_{n-1} & \cdots & a_0 & & & \\ & & a_n & a_{n-1} & \cdots & a_0 & & \\ & & & \ddots & & & & \\ & & & & a_n & a_{n-1} & \cdots & a_0 \\ b_m & b_{m-1} & \cdots & b_0 & & & & \\ & b_m & b_{m-1} & \cdots & b_0 & & & \\ & & b_m & b_{m-1} & \cdots & b_0 & & \\ & & & \ddots & & & & \\ & & & & b_m & b_{m-1} & \cdots & b_0 \end{vmatrix}$$

is zero. Here $R(f,g)$, called the *resultant* of the two polynomials, is the determinant of an $(n+m) \times (n+m)$ matrix R with m rows involving the coefficients of $f(x)$ and n rows involving the coefficients of $g(x)$.

30. **(a)** With notations as in the previous problem, show that we have the matrix equation

$$R\begin{pmatrix} x^{n+m-1} \\ x^{n+m-2} \\ \vdots \\ x \\ 1 \end{pmatrix} = \begin{pmatrix} x^{m-1}f(x) \\ x^{m-2}f(x) \\ \vdots \\ f(x) \\ x^{n-1}g(x) \\ x^{n-2}g(x) \\ \vdots \\ g(x) \end{pmatrix}.$$

(b) Let R' denote the matrix of cofactors of R as in Theorem 30 of Section 11.4, so $R'R = R(f,g)I$, where I is the identity matrix. Multiply both sides of the matrix equation above by R' and equate the bottom entry of the resulting column matrices to prove that there are polynomials $r(x), s(x) \in F[x]$ such that $R(f,g)$ is equal to $r(x)f(x) + s(x)g(x)$, i.e., the resultant of two polynomials is a linear combination (in $F[x]$) of the polynomials.

31. Consider $f(x)$ and $g(x)$ as general polynomials and suppose the roots of $f(x)$ are $x_1, \ldots, x_n$ and the roots of $g(x)$ are $y_1, \ldots, y_m$. The coefficients of $f(x)$ are powers of a_n times the elementary symmetric functions in $x_1, x_2, \ldots, x_n$ and the coefficients of $g(x)$ are powers of b_m times the elementary symmetric functions in $y_1, y_2, \ldots, y_m$.

(a) By expanding the determinant show that $R(f,g)$ is homogeneous of degree m in the coefficients a_i and homogeneous of degree n in the coefficients b_j.

(b) Show that $R(f,g)$ is $a_n^m b_m^n$ times a symmetric function in $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$.

(c) Since $R(f,g)$ is 0 if $f(x)$ and $g(x)$ have a common root, say $x_i = y_j$, show that $R(f,g)$ is divisible by $x_i - y_j$ for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Conclude by degree considerations that

$$R = a_n^m b_m^n \prod_{i=1}^{n} \prod_{j=1}^{m} (x_i - y_j).$$

(d) Show that the product in (c) can be also be written

$$R(f, g) = a_n^m \prod_{i=1}^{n} g(x_i) = (-1)^{nm} b_m^n \prod_{j=1}^{m} f(y_j).$$

This gives an interesting ***reciprocity*** between the product of g evaluated at the roots of f and the product of f evaluated at the roots of g.

32. Consider now the special case where $g(x) = f'(x)$ is the derivative of the polynomial $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ and suppose the roots of $f(x)$ are $\alpha_1, \alpha_2, \ldots, \alpha_n$. Using the formula

$$R(f, f') = \prod_{i=1}^{n} f'(\alpha_i)$$

of the previous exercise, prove that

$$D = (-1)^{n(n-1)/2} R(f, f')$$

where D is the discriminant of $f(x)$.

33. **(a)** Prove that the discriminant of the cyclotomic polynomial $\Phi_p(x)$ of the p^{th} roots of unity for an odd prime p is $(-1)^{(p-1)/2} p^{p-2}$ [One approach: use Exercise 5 of the previous section together with the determinant form for the discriminant in terms of the power sums p_i.]

(b) Prove that $\mathbb{Q}(\sqrt{(-1)^{(p-1)/2} p}) \subset \mathbb{Q}(\zeta_p)$ for p an odd prime. (Cf. also Exercise 11 of Section 7.)

34. Use the previous exercise to prove that every quadratic extension of $\mathbb{Q}$ is contained in a cyclotomic extension (a special case of the Kronecker–Weber Theorem).

35. Prove that the discriminant D of the polynomial $x^n + px + q$ is given by the formula $(-1)^{n(n-1)/2} n^n q^{n-1} + (-1)^{(n-1)(n-2)/2} (n-1)^{n-1} p^n$.

36. Prove that the discriminant of $x^n + nx^{n-1} + n(n-1)x^{n-2} + \cdots + n(n-1)\ldots(3)(2)x + n!$ is $(-1)^{n(n-1)/2}(n!)^n$.

The following exercises 37 to 43 outline two procedures for writing a symmetric function in terms of the elementary symmetric functions. Let $f(x_1, \ldots, x_n)$ be a polynomial which is symmetric in $x_1, \ldots, x_n$. Recall that the degree (sometimes called the *weight*) of the monomial $Ax_1^{a_1} x_2^{a_2} \ldots x_n^{a_n}$ $(a_i \geq 0)$ is $a_1 + a_2 + \cdots + a_n$ and that a polynomial is *homogeneous (of degree m)* if every monomial has the same degree (m).

37. **(a)** Show that every polynomial $f(x_1, \ldots, x_n)$ can be written as a sum of homogeneous polynomials. Show that if $f(x_1, \ldots, x_n)$ is symmetric then each of these homogeneous polynomials is also symmetric.

(b) Show that the monomial $Bs_1^{a_1} s_2^{a_2} \ldots s_n^{a_n}$ in the elementary symmetric functions is a homogeneous polynomial in $x_1, x_2, \ldots, x_n$ of degree $a_1 + 2a_2 + \cdots + na_n$.

In writing $f(x_1, \ldots, x_n)$ as a polynomial in the symmetric functions it therefore suffices to assume that $f(x_1, \ldots, x_n)$ is homogeneous.

Recall the *lexicographic monomial order* with $x_1 > x_2 > \cdots > x_n$ defined in Section 9.6, where the nonzero monomial term with exponents $(a_1, a_2, \ldots, a_n)$ comes before the nonzero monomial term with exponents $(b_1, b_2, \ldots, b_n)$ if the initial components of the two n-tuples of exponents are equal and the first component where they differ has $a_i > b_i$. If $f(x_1, \ldots, x_n)$ contains the monomial $Ax_1^{a_1}x_2^{a_2}\ldots x_n^{a_n}$ then since $f(x_1, \ldots, x_n)$ is symmetric it also contains all the permuted monomials. Among these choose the lexicographically largest monomial, which therefore satisfies $a_1 \geq a_2 \geq \cdots \geq a_n \geq 0$.

38. **(a)** Show that the monomial $As_1^{a_1-a_2}s_2^{a_2-a_3}\cdots s_n^{a_n}$ in the elementary symmetric functions has the same lexicographic initial term.

(b) Show that subtracting $As_1^{a_1-a_2}s_2^{a_2-a_3}\cdots s_n^{a_n}$ from $f(x)$ yields either 0 or a symmetric polynomial of the same degree whose terms are lexicographically smaller than the terms in $f(x_1, \ldots, x_n)$.

(c) Show that the iteration of this procedure (lexicographic ordering, choosing the lexicographically largest term, subtracting the associated monomial in the elementary symmetric functions) terminates, expressing $f(x_1, \ldots, x_n)$ as a polynomial in the elementary symmetric functions.

39. Use the algorithm described in Exercise 38 to prove that a polynomial $f(x_1, \ldots, x_n)$ that is symmetric in $x_1, \ldots, x_n$ can be expressed *uniquely* as a polynomial in the elementary symmetric functions.

40. Use the procedure in Exercise 38 to express each of the following symmetric functions as a polynomial in the elementary symmetric functions:

(a) $(x_1 - x_2)^2$

(b) $x_1^2 + x_2^2 + x_3^2$

(c) $x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2$.

41. Use the procedure in Exercise 38 to express $\sum_{i \neq j} x_i^2 x_j$ as a polynomial in the elementary symmetric functions.

We now know that a symmetric polynomial $f(x_1, \ldots, x_n)$ can be written uniquely as a polynomial in the elementary symmetric functions. Using this existence and uniqueness we can describe an alternate and computationally useful method for determining the coefficients of the elementary symmetric functions in this polynomial. As in Exercise 37 we may assume that $f(x_1, \ldots, x_n)$ is homogeneous of degree M. Let N be the maximum degree of any of the variables $x_1, \ldots, x_n$ in $f(x_1, \ldots, x_n)$.

(a) Determine all of the possible monomials $A_i s_1^{a_1} s_2^{a_2} \cdots s_n^{a_n}$ appearing in $f(x_1, \ldots, x_n)$ from the constraints

$$\begin{aligned} a_1 + 2a_2 + \cdots + na_n &= M \\ a_1 + a_2 + \cdots + a_n &\leq N. \end{aligned}$$

(b) Since $f(x_1, \ldots, x_n) = \sum A_i s_1^{a_1} s_2^{a_2} \cdots s_n^{a_n}$ is a polynomial *identity*, it is valid for any substitution of values for $x_1, \ldots, x_n$. Each substitution into this equation gives a linear relation on the coefficients A_i and so a sufficient number of substitutions will determine the A_i.

42. Show that the function $(x_1 + x_2 - x_3 - x_4)(x_1 + x_3 - x_2 - x_4)(x_1 + x_4 - x_2 - x_3)$ is symmetric in x_1, x_2, x_3, x_4 and use the preceding procedure to prove it can be expressed as a polynomial in the elementary symmetric functions as $s_1^3 - 4s_1s_2 + 8s_3$.

43. Express each of the following in terms of the elementary symmetric functions:

(a) $\sum_{i \neq j} x_i^2 x_j$ **(b)** $\sum_{i,j,k \text{ distinct}} x_i^2 x_j x_k$ **(c)** $\sum_{i,j,k \text{ distinct}} x_i^2 x_j^2 x_k^2$.

44. Let $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ be the roots of a quartic polynomial $f(x)$ over $\mathbb{Q}$. Show that the quantities $\alpha_1\alpha_2 + \alpha_3\alpha_4$, $\alpha_1\alpha_3 + \alpha_2\alpha_4$, and $\alpha_1\alpha_4 + \alpha_2\alpha_3$ are permuted by the Galois group of $f(x)$. Conclude that these elements are the roots of a cubic polynomial with coefficients in $\mathbb{Q}$ (also sometimes referred to as the *resolvent cubic* of $f(x)$).

45. If $f(x) = x^3 + px + q \in \mathbb{Z}[x]$ is irreducible, prove that its discriminant $D = -4p^3 - 27q^2$ is an integer not equal to 0, ±1.

46. Prove that every finite group occurs as the Galois group of a field extension of the form $F(x_1, x_2, \ldots, x_n)/E$.

47. Let F be a field of characteristic 0 in which every cubic polynomial has a root. Let $f(x)$ be an irreducible quartic polynomial over F whose discriminant is a square in F. Determine the Galois group of $f(x)$.

48. This exercise determines the splitting field K for the polynomial $f(x) = x^6 - 2x^3 - 2$ over $\mathbb{Q}$ (cf. also Exercise 2 of Section 8).

(a) Prove that $f(x)$ is irreducible over $\mathbb{Q}$ with roots the three cube roots of $1 \pm \sqrt{3}$.

(b) Prove that K contains the field $\mathbb{Q}(\sqrt{-3})$ of 3rd roots of unity and contains $\mathbb{Q}(\sqrt{3})$, hence contains the biquadratic field $F = \mathbb{Q}(i, \sqrt{3})$. Take the product of two of the roots in (a) to prove that K contains $\sqrt[3]{2}$ and conclude that K is an extension of the field $L = \mathbb{Q}(\sqrt[3]{2}, i, \sqrt{3})$.

(c) Prove that $[L : \mathbb{Q}] = 12$ and that K is obtained from L by adjoining the cube root of an element in L, so that $[K : \mathbb{Q}] = 12$ or 36.

(d) Prove that if $[K : \mathbb{Q}] = 12$ then $K = \mathbb{Q}(\sqrt[3]{2}, i, \sqrt{3})$ and that $\mathrm{Gal}(K/\mathbb{Q})$ is isomorphic to the direct product of the cyclic group of order 2 and S_3. Prove that if $[K : \mathbb{Q}] = 12$ then there is a unique real cubic subfield in K, namely $\mathbb{Q}(\sqrt[3]{2})$.

(e) Take the quotient of the two real roots in (a) to show that $\sqrt[3]{2 + \sqrt{3}}$ and $\sqrt[3]{2 - \sqrt{3}}$ (real roots) are both elements of K. Show that $\alpha = \sqrt[3]{2 + \sqrt{3}} + \sqrt[3]{2 - \sqrt{3}}$ is a real root of the irreducible cubic equation $x^3 - 3x - 4$ whose discriminant is $-2^2 3^4$. Conclude that the Galois closure of $\mathbb{Q}(\alpha)$ contains $\mathbb{Q}(i)$ so in particular $\mathbb{Q}(\alpha) \neq \mathbb{Q}(\sqrt[3]{2})$.

(f) Conclude from (e) that $G = \mathrm{Gal}(K/\mathbb{Q})$ is of order 36. Determine all the elements of G explicitly and in particular show that G is isomorphic to $S_3 \times S_3$.

49. Prove that the Galois group over $\mathbb{Q}$ of $x^6 - 4x^3 + 1$ is isomorphic to the dihedral group of order 12. [Observe that the two real roots are inverses of each other.]

50. *(Criterion for the Galois Group of an Irreducible Cubic over an Arbitrary Field)* Suppose K is a field and $f(x) = x^3 + ax^2 + bx + c \in K[x]$ is irreducible, so the Galois group of $f(x)$ over K is either S_3 or A_3.

(a) Show that the Galois group of $f(x)$ is A_3 if and only if the resultant quadratic polynomial $g(x) = x^2 + (ab - 3c)x + (b^3 + a^3c - 6abc + 9c^2)$ has a root in K. [If α, β, γ are the roots of $f(x)$ show that the Galois group is A_3 if and only if the element $\theta = \alpha\beta^2 + \beta\gamma^2 + \gamma\alpha^2$ is an element of K and that θ is a root of $g(x)$.] Show that the discriminant of $g(x)$ is the same as the discriminant of $f(x)$.

(b) ($\mathrm{ch}(K) \neq 2$) If K has characteristic different from 2 show either from (a) or directly from the definition of the discriminant that the Galois group of $f(x)$ is A_3 if and only if the discriminant of $f(x)$ is a square in K.

(c) ($\mathrm{ch}(K) = 2$) If K has characteristic 2 show that the discriminant of $f(x)$ is always a square. Show that $f(x)$ can be taken to be of the form $x^3 + px + q$ and that the Galois group of $f(x)$ is A_3 if and only if the quadratic $x^2 + qx + (p^3 + q^2)$ has a root in K (equivalently, if $(p^3 + q^2)/q^2 \in K$ is in the image of the *Artin–Schreier map* $x \mapsto x^2 - x$ mapping K to K).

(d) If $K = \mathbb{F}_2(t)$ where t is transcendental over $\mathbb{F}_2$. Prove that the polynomials $x^3 + t^2x + t^3$, $x^3 + (t^2 + t + 1)x + (t^2 + t + 1)$, and $x^3 + (t^2 + t + 1)x + (t^3 + t^2 + t)$ have A_3 as Galois group while $x^3 + t^2x + t$ and $x^3 = x + t$ have S_3 as Galois group.

51. This exercise proves *Sturm's Theorem* determining the number of real roots of a polynomial $f(x) \in \mathbb{R}[x]$ in an interval $[a, b]$. The multiple roots of $f(x)$ are zeros of the g.c.d. of $f(x)$ and its derivative $f'(x)$, and it follows that to determine the real roots of $f(x)$ in $[a, b]$ we may assume that the roots of $f(x)$ are *simple*.

Apply the Euclidean algorithm to $f_0(x) = f(x)$ and its derivative $f_1(x) = f'(x)$ using the *negative* of the remainder at each stage to find a sequence of polynomials $f(x), f'(x), f_2(x), \dots, f_n(x)$ with

$$f_{i-1}(x) = q_i(x)f_i(x) - f_{i+1}(x) \qquad i = 0, 1, \dots, n-1$$

where $f_n(x) \in \mathbb{R}$ is a nonzero constant.

(a) Prove that consecutive polynomials $f_i(x), f_{i+1}(x)$ for $i = 0, 1, \dots, n-1$ have no common zeros. [Show that otherwise $f_{i+2}(c) = f_{i+3}(c) = \cdots = 0$, and derive a contradiction.]

(b) If $f_i(c) = 0$ for some $i = 0, 1, \dots, n-1$, prove that one of the two values $f_{i-1}(c)$, $f_{i+1}(c)$ is strictly negative and the other is strictly positive.

For any real number α, let $V(\alpha)$ denote the number of sign changes in the *Sturm sequence* of real numbers

$$f(\alpha), f'(\alpha), f_2(\alpha), \dots, f_n(\alpha),$$

ignoring any 0's that appear (for example $-1, -2, 0, +3, -4$ has signs $--+-$ disregarding the 0, so there are 2 sign changes, the first from -2 to $+3$, the second from $+3$ to -4).

(c) Suppose $\alpha < \beta$ and that all the elements in the Sturm sequences for α and for β are nonzero. Prove that unless $f_i(c) = 0$ for some $\alpha < c < \beta$ and some $i = 0, 1, \dots, n-1$, then the signs of all the elements in these two Sturm sequences are the same, so in particular $V(\alpha) = V(\beta)$.

(d) If $f_j(c) = 0$ prove that there is a sufficiently small interval (α, β) containing c so that $f_j(x)$ has no zero other than c for $\alpha < x < \beta$.

(e) If $j \geq 1$ in (d), prove that the number of sign changes in $f_{j-1}(\alpha), f_j(\alpha), f_{j+1}(\alpha)$ and in $f_{j-1}(\beta), f_j(\beta), f_{j+1}(\beta)$ are the same. [Observe that $f_{j-1}(c)$ and $f_{j+1}(c)$ have opposite signs by (b) and $f_{j-1}(x)$ and $f_{j+1}(x)$ do not change sign in (α, β).]

(f) If $j = 0$ in (d) show that the number of sign changes in $f(\alpha), f'(\alpha)$ is one more than the number of sign changes in $f(\beta), f'(\beta)$. [If $f'(c) > 0$ then $f(x)$ is increasing at c, so that $f(\alpha) < 0$, $f(\beta) > 0$, and $f'(x)$ does not change sign in (α, β), so the signs change from $-+$ to $++$. Similarly if $f'(c) < 0$.]

(g) Prove *Sturm's Theorem*: if $f(x)$ is a polynomial with real coefficients all of whose real roots are simple then the number of real zeros of $f(x)$ in an interval $[a, b]$ where $f(a)$ and $f(b)$ are both nonzero is given by $V(a) - V(b)$. [Use (c), (e) and (f) to see that as α runs from a to b the number $V(\alpha)$ of sign changes is constant unless α passes through a zero of $f(x)$, in which case it decreases by precisely 1.]

(h) Suppose $f(x) = x^5 + px + q \in \mathbb{R}[x]$ has simple roots. Show that the sequence of polynomials above is given by $f(x)$, $5x^4 + p$, $(-4p/5)x + q$, and $-D/(256p^4)$ where $D = 256p^5 + 3125q^4$ is the discriminant of $f(x)$. Conclude for $p > 0$ that $f(x)$ has precisely one real root and for $p < 0$ that $f(x)$ has precisely 1 or 3 real roots depending on whether $D > 0$ or $D < 0$, respectively. [E.g., if $p < 0$ and $D < 0$ then at $-\infty$ the signs are $- + -+$ with 3 sign changes and at $+\infty$ the signs are $+ + ++$ with no sign changes.]

14.7 SOLVABLE AND RADICAL EXTENSIONS: INSOLVABILITY OF THE QUINTIC

We now investigate the question of solving for the roots of a polynomial by *radicals*, that is, in terms of the algebraic operations of addition, subtraction, multiplication, division and the extraction of n^{th} roots. The quadratic formula for the roots of a polynomial of degree 2 is familiar from elementary algebra and we shall derive below similar formulas for the roots of cubic and quartic polynomials. For polynomials of degree ≥ 5, however, we shall see that such formulas are not possible — this is Abel's Theorem on the insolvability of the general quintic. The reason for this is quite simple: we shall see that a polynomial is solvable by radicals if and only if its Galois group is a solvable group (which explains the terminology) and for $n \geq 5$ the group S_n is not solvable.

We first discuss *simple radical extensions*, namely extensions obtained by adjoining to a field F the n^{th} root of an element a in F. Since all the roots of the polynomial $x^n - a$ for $a \in F$ differ by factors of the n^{th} roots of unity, adjoining one such root will give a Galois extension if and only if this field contains the n^{th} roots of unity. Simple radical extensions are best behaved when the base field F already contains the appropriate roots of unity. The symbol $\sqrt[n]{a}$ for $a \in F$ will be used to denote any root of the polynomial $x^n - a \in F[x]$.

Definition. The extension K/F is said to be *cyclic* if it is Galois with a cyclic Galois group.

Proposition 36. Let F be a field of characteristic not dividing n which contains the n^{th} roots of unity. Then the extension $F(\sqrt[n]{a})$ for $a \in F$ is cyclic over F of degree dividing n.

Proof: The extension $K = F(\sqrt[n]{a})$ is Galois over F if F contains the n^{th} roots of unity since it is the splitting field for $x^n - a$. For any $\sigma \in \text{Gal}(K/F)$, $\sigma(\sqrt[n]{a})$ is another root of this polynomial, hence $\sigma(\sqrt[n]{a}) = \zeta_\sigma \sqrt[n]{a}$ for some n^{th} root of unity ζ_σ. This gives a map

$$\begin{aligned} \text{Gal}(K/F) &\to \mu_n \\ \sigma &\mapsto \zeta_\sigma \end{aligned}$$

where μ_n denotes the group of n^{th} roots of unity. Since F contains μ_n, every n^{th} root of unity is fixed by every element of $\text{Gal}(K/F)$. Hence

$$\begin{aligned} \sigma\tau(\sqrt[n]{a}) &= \sigma(\zeta_\tau \sqrt[n]{a}) \\ &= \zeta_\tau \sigma(\sqrt[n]{a}) \\ &= \zeta_\tau \zeta_\sigma \sqrt[n]{a} = \zeta_\sigma \zeta_\tau \sqrt[n]{a} \end{aligned}$$

which shows that $\zeta_{\sigma\tau} = \zeta_\sigma \zeta_\tau$, so the map above is a homomorphism. The kernel consists precisely of the automorphisms which fix $\sqrt[n]{a}$, namely the identity. This gives an injection of $\text{Gal}(K/F)$ into the cyclic group μ_n of order n, which proves the proposition.

Let now K be any cyclic extension of degree n over a field F of characteristic not dividing n which contains the n^{th} roots of unity. Let σ be a generator for the cyclic group $\text{Gal}(K/F)$.

Definition. For $\alpha \in K$ and any n^{th} root of unity ζ, define the *Lagrange resolvent* $(\alpha, \zeta) \in K$ by

$$(\alpha, \zeta) = \alpha + \zeta\sigma(\alpha) + \zeta^2\sigma^2(\alpha) + \cdots + \zeta^{n-1}\sigma^{n-1}(\alpha).$$

If we apply the automorphism σ to (α, ζ) we obtain

$$\sigma(\alpha, \zeta) = \sigma\alpha + \zeta\sigma^2(\alpha) + \zeta^2\sigma^3(\alpha) + \cdots + \zeta^{n-1}\sigma^n(\alpha)$$

since ζ is an element of the base field F so is fixed by σ. We have $\zeta^n = 1$ in μ_n and $\sigma^n = 1$ in $\text{Gal}(K/F)$ so this can be written

$$\begin{aligned}\sigma(\alpha, \zeta) &= \sigma\alpha + \zeta\sigma^2(\alpha) + \zeta^2\sigma^3(\alpha) + \cdots + \zeta^{-1}\alpha \\ &= \zeta^{-1}(\alpha + \zeta\sigma(\alpha) + \zeta^2\sigma^2(\alpha) + \cdots + \zeta^{n-1}\sigma^{n-1}(\alpha)) \\ &= \zeta^{-1}(\alpha, \zeta). \end{aligned} \qquad (14.19)$$

It follows that

$$\sigma(\alpha, \zeta)^n = (\zeta^{-1})^n(\alpha, \zeta)^n = (\alpha, \zeta)^n$$

so that $(\alpha, \zeta)^n$ is fixed by $\text{Gal}(K/F)$, hence is an element of F for any $\alpha \in K$.

Let ζ be a primitive n^{th} root of unity. By the linear independence of the automorphisms $1, \sigma, \ldots, \sigma^{n-1}$ (Theorem 7), there is an element $\alpha \in K$ with $(\alpha, \zeta) \neq 0$. Iterating (19) we have

$$\sigma^i(\alpha, \zeta) = \zeta^{-i}(\alpha, \zeta), \qquad i = 0, 1, \ldots,$$

and it follows that σ^i does not fix (α, ζ) for any $i < n$. Hence this element cannot lie in any proper subfield of K, so $K = F((\alpha, \zeta))$. Since we proved $(\alpha, \zeta)^n = a \in F$ above, we have $F(\sqrt[n]{a}) = F((\alpha, \zeta)) = K$. This proves the following converse of Proposition 36.

Proposition 37. Any cyclic extension of degree n over a field F of characteristic not dividing n which contains the n^{th} roots of unity is of the form $F(\sqrt[n]{a})$ for some $a \in F$.

Remark: The two propositions above form a part of what is referred to as *Kummer theory*. A group G is said to have *exponent n* if $g^n = 1$ for every $g \in G$. Let F be a field of characteristic not dividing n which contains the n^{th} roots of unity. If we take elements $a_1, \ldots, a_k \in F^\times$ then as in Proposition 36 we can see that the extension

$$F(\sqrt[n]{a_1}, \sqrt[n]{a_2}, \ldots, \sqrt[n]{a_k}) \qquad (14.20)$$

is an abelian extension of F whose Galois group is of exponent n. Conversely, any abelian extension of exponent n is of this form.

Denote by $(F^\times)^n$ the subgroup of the multiplicative group $F^\times$ consisting of the n^{th} powers of nonzero elements of F. The quotient group $F^\times/(F^\times)^n$ is an abelian group of exponent n. The Galois group of the extension in (20) is isomorphic to the group generated in $F^\times/(F^\times)^n$ by the elements $a_1, \ldots, a_k$ and two extensions as in (20) are equal if and only if their associated groups in $F^\times/(F^\times)^n$ are equal.

Hence the (finitely generated) subgroups of $F^\times/(F^\times)^n$ classify the abelian extensions of exponent n over fields containing the n^{th} roots of unity (and characteristic not dividing n). Such extensions are called *Kummer extensions.*

These results generalize the case $k = 1$ above and can be proved in a similar way.

For simplicity we now consider the situation of a base field F of characteristic 0. As in the previous propositions the results are valid over fields whose characteristics do not divide any of the orders of the roots that will be taken.

Definition.

(1) An element α which is algebraic over F can be *expressed by radicals* or *solved for in terms of radicals* if α is an element of a field K which can be obtained by a succession of simple radical extensions
$$F = K_0 \subset K_1 \subset \cdots \subset K_i \subset K_{i+1} \subset \cdots \subset K_s = K \tag{14.21}$$
where $K_{i+1} = K_i(\sqrt[n_i]{a_i})$ for some $a_i \in K_i$, $i = 0, 1, \dots, s-1$. Here $\sqrt[n_i]{a_i}$ denotes some root of the polynomial $x^{n_i} - a_i$. Such a field K will be called a *root extension* of F.

(2) A polynomial $f(x) \in F[x]$ can be *solved by radicals* if all its roots can be solved for in terms of radicals.

This gives a precise meaning to the intuitive notion that α is obtained by successive algebraic operations (addition, subtraction, multiplication and division) and successive root extractions. For example, the element
$$-1+\sqrt{17}+\sqrt{2(17-\sqrt{17})}+2\sqrt{17+3\sqrt{17}-\sqrt{2(17-\sqrt{17})}-2\sqrt{2(17+\sqrt{17})}}$$
encountered at the end of Section 5 (used to construct the regular 17-gon) is expressed by radicals and is contained in the field K_4, where
$$\begin{aligned}
&K_0 = \mathbb{Q} \\
&K_1 = K_0(\sqrt{a_0}) && a_0 = 17 \\
&K_2 = K_1(\sqrt{a_1}) && a_1 = 2(17-\sqrt{17}) \\
&K_3 = K_2(\sqrt{a_2}) && a_2 = 2(17+\sqrt{17}) \\
&K_4 = K_3(\sqrt{a_3}) && a_3 = 17+3\sqrt{17}-\sqrt{2(17-\sqrt{17})}-2\sqrt{2(17+\sqrt{17})}.
\end{aligned}$$
Each of these extensions is a radical extension. The fact that no roots other than square roots are required reflects the fact that the regular 17-gon is constructible by straightedge and compass.

In considering radical extensions one may always adjoin roots of unity, since by definition the roots of unity are radicals. This is useful because then cyclic extensions become radical extensions and conversely. In particular we have:

Lemma 38. If α is contained in a root extension K as in (21) above, then α is contained in a root extension which is Galois over F and where each extension K_{i+1}/K_i is cyclic.

Proof: Let L be the Galois closure of K over F. For any $\sigma \in \text{Gal}(L/F)$ we have the chain of subfields

$$F = \sigma K_0 \subset \sigma K_1 \subset \cdots \subset \sigma K_i \subset \sigma K_{i+1} \subset \cdots \subset \sigma K_s = \sigma K$$

where $\sigma K_{i+1}/\sigma K_i$ is again a simple radical extension (since it is generated by the element $\sigma(\sqrt[n_i]{a_i})$, which is a root of the equation $x^{n_i} - \sigma(a_i)$ over $\sigma(K_i)$). It is easy to see that the composite of two root extensions is again a root extension (if K' is another root extension with subfields K_i', first take the composite of K_1' with the fields $K_0, K_1, \ldots, K_s$, then the composite of these fields with K_2', etc. so that each individual extension in this process is a simple radical extension). It follows that the composite of all the conjugate fields $\sigma(K)$ for $\sigma \in \text{Gal}(L/F)$ is again a root extension. Since this field is precisely L, we see that α is contained in a Galois root extension.

We now adjoin to F the n_i-th roots of unity for all the roots $\sqrt[n_i]{a_i}$ of the simple radical extensions in the Galois root extension K/F, obtaining the field F', say, and then form the composite of F' with the root extension:

$$F \subseteq F' = F'K_0 \subseteq F'K_1 \subseteq \cdots \subseteq F'K_i \subseteq F'K_{i+1} \subseteq \cdots \subseteq F'K_s = F'K.$$

The field $F'K$ is a Galois extension of F since it is the composite of two Galois extensions. The extension from F to $F' = F'K_0$ can be given as a chain of subfields with each individual extension cyclic (this is true for any abelian extension). Each extension $F'K_{i+1}/F'K_i$ is a simple radical extension and since we now have the appropriate roots of unity in the base fields, each of these individual extensions from F' to $F'K$ is a cyclic extension by Proposition 36. Hence $F'K/F$ is a root extension which is Galois over F with cyclic intermediate extensions, completing the proof.

Recall from Section 3.4 (cf. also Section 6.1) that a finite group G is *solvable* if there exists a chain of subgroups

$$1 = G_s \le G_{s-1} \le \cdots \le G_{i+1} \le G_i \le \cdots \le G_0 = G \qquad (14.22)$$

with G_i/G_{i+1} cyclic, $i = 0, 1, \ldots, s-1$. We have proved that subgroups and quotient groups of solvable groups are solvable and that if $H \le G$ and G/H are both solvable, then G is solvable.

We now prove Galois' fundamental connection between solving for the roots of polynomials in terms of radicals and the Galois group of the polynomial. We continue to work over a field F of characteristic 0, but it is easy to see that the proof is valid over any field of characteristic not dividing the order of the Galois group or the orders of the radicals involved.

Theorem 39. The polynomial $f(x)$ can be solved by radicals if and only if its Galois group is a solvable group.

Proof: Suppose first that $f(x)$ can be solved by radicals. Then each root of $f(x)$ is contained in an extension as in the lemma. The composite L of such extensions is again of the same type by Proposition 21. Let G_i be the subgroups corresponding to the subfields K_i, $i = 0, 1, \dots, s-1$. Since

$$\mathrm{Gal}(K_{i+1}/K_i) = G_i/G_{i+1} \qquad i = 0, 1, \dots, s-1$$

it follows that the Galois group $G = \mathrm{Gal}(L/F)$ is a solvable group. The field L contains the splitting field of $f(x)$ so the Galois group of $f(x)$ is a quotient group of the solvable group G, hence is solvable.

Suppose now that the Galois group G of $f(x)$ is a solvable group and let K be the splitting field for $f(x)$. Taking the fixed fields of the subgroups in a chain (22) for G gives a chain

$$F = K_0 \subset K_1 \subset \cdots \subset K_i \subset K_{i+1} \subset \cdots \subset K_s = K$$

where K_{i+1}/K_i, $i = 0, 1, \dots, s-1$ is a cyclic extension of degree n_i. Let F' be the cyclotomic field over F of all roots of unity of order n_i, $i = 0, 1, \dots, s-1$ and form the composite fields $K_i' = F'K_i$. We obtain a sequence of extensions

$$F \subseteq F' = F'K_0 \subseteq F'K_1 \subseteq \cdots \subseteq F'K_i \subseteq F'K_{i+1} \subseteq \cdots \subseteq F'K_s = F'K.$$

The extension $F'K_{i+1}/F'K_i$ is cyclic of degree dividing n_i, $i = 0, 1, \dots, s-1$ (by Proposition 19). Since we now have the appropriate roots of unity in the base fields, each of these cyclic extensions is a simple radical extension by Proposition 37. Each of the roots of $f(x)$ is therefore contained in the root extension $F'K$ so that $f(x)$ can be solved by radicals.

Corollary 40. The general equation of degree n cannot be solved by radicals for $n \geq 5$.

Proof: For $n \geq 5$ the group S_n is not solvable as we showed in Chapter 4. The corollary follows immediately from Theorems 32 and 39.

This corollary shows that there is no formula involving radicals analogous to the quadratic formula for polynomials of degree 2 for the roots of a polynomial of degree 5. To give an example of a *specific* polynomial over $\mathbb{Q}$ of degree 5 whose roots cannot be expressed in terms of radicals we must demonstrate a polynomial of degree 5 with rational coefficients having S_5 (or A_5, which is also not solvable) as Galois group (cf. also Exercise 21, which gives a criterion for the solvability of a quintic).

Example

Consider the polynomial $f(x) = x^5 - 6x + 3 \in \mathbb{Q}[x]$. This polynomial is irreducible since it is Eisenstein at 3. The splitting field K for this polynomial therefore has degree divisible by 5, since adjoining one root of $f(x)$ to $\mathbb{Q}$ generates an extension of degree 5. The Galois group G is therefore a subgroup of S_5 of order divisible by 5 so contains an element of order 5. The only elements in S_5 of order 5 are 5-cycles, so G contains a 5-cycle.

Since $f(-2) = -17$, $f(0) = 3$, $f(1) = -2$, and $f(2) = 23$ we see that $f(x)$ has a real root in each of the intervals $(-2, 0)$, $(0, 1)$ and $(1, 2)$. By the Mean Value Theorem, if there were 4 real roots then the derivative $f'(x) = 5x^4 - 6$ would have at least 3 real zeros, which it does not. Hence these are the only real roots. (This also follows easily by Descartes' rule of signs.) By the Fundamental Theorem of Algebra $f(x)$ has 5 roots in $\mathbb{C}$. Hence $f(x)$ has two complex roots which are not real. Let τ denote the automorphism of complex conjugation in $\mathbb{C}$. Since the coefficients of $f(x)$ are real, the two complex roots must be interchanged by τ (since they are not fixed, not being real). Hence the restriction of complex conjugation to K fixes three of the roots of $f(x)$ and interchanges the other two. As an element of G, $\tau|_K$ is therefore a transposition.

It is now a simple exercise to show that any 5-cycle together with any transposition generate all of S_5. It follows that $G = S_5$, so the roots of $x^5 - 6x + 3$ cannot be expressed by radicals.

As indicated in this example, a great deal of information regarding the Galois group can be obtained by understanding the *cycle types* of the automorphisms in G considered as a subgroup of S_n. In practice this is the most efficient way of determining the Galois groups of polynomials of degrees ≥ 5 (becoming more difficult the larger the degree, of course, if only because the possible subgroups of S_n are vastly more numerous). We describe this procedure in the next section.

By Theorem 39, any polynomial of degree $n \leq 4$ can be solved by radicals, since S_n is a solvable group for these n. For $n = 2$ this is just the familiar quadratic formula. For $n = 3$ the formula is known as *Cardano's Formula* (named for Geronimo Cardano (1501–1576)) and the formula for $n = 4$ can be reduced to this one. The formulas are valid over any field F of characteristic $\neq 2, 3$, which are the characteristics dividing the orders of the radicals necessary and the orders of the possible Galois groups (which are subgroups of S_3 and S_4). For simplicity we shall derive the formulas over $\mathbb{Q}$.

Solution of Cubic Equations by Radicals: Cardano's Formulas

From the proof of Theorem 39 and the fact that a composition series for S_3 as in equation (22) is given by $1 \leq A_3 \leq S_3$ we should expect that the solution of the cubic

$$f(x) = x^3 + ax^2 + bx + c$$

(or equivalently, under the substitution $x = y - a/3$,

$$g(y) = y^3 + py + q,$$

where

$$p = \frac{1}{3}(3b - a^2) \qquad q = \frac{1}{27}(2a^3 - 9ab + 27c)\ )$$

to involve adjoining the 3rd roots of unity and the formation of Lagrange resolvents involving these roots of unity.

Let ρ denote a primitive 3rd root of unity, so that $\rho^2 + \rho + 1 = 0$. Let the roots of $g(y)$ be α, β, and γ, so that

$$\alpha + \beta + \gamma = 0$$

(one of the reasons for changing from $f(x)$ to $g(x)$). Over the field $\mathbb{Q}(\sqrt{D})$ where D is the discriminant (computed in the last section) the Galois group of $g(y)$ is A_3, i.e., a cyclic group of order 3. If we adjoin ρ then this extension is a radical extension of degree 3, with generator given by a Lagrange Resolvent, as in the proof of Proposition 37. Consider therefore the elements

$$\begin{aligned}(\alpha, 1) &= \alpha + \beta + \gamma = 0\\ \theta_1 = (\alpha, \rho) &= \alpha + \rho\beta + \rho^2\gamma\\ \theta_2 = (\alpha, \rho^2) &= \alpha + \rho^2\beta + \rho\gamma.\end{aligned}$$

Note that the sum of these resolvents is

$$\theta_1 + \theta_2 = 3\alpha \tag{14.23}$$

since $1 + \rho + \rho^2 = 0$. Similarly

$$\begin{aligned}\rho^2\theta_1 + \rho\theta_2 &= 3\beta\\ \rho\theta_1 + \rho^2\theta_2 &= 3\gamma.\end{aligned} \tag{14.23$'$}$$

We also showed in general before Proposition 37 that the cube of these resolvents must lie in $\mathbb{Q}(\sqrt{D}, \rho)$. Expanding θ_1^3 we obtain

$$\begin{aligned}\alpha^3 + \beta^3 + \gamma^3 &+ 3\rho(\alpha^2\beta + \beta^2\gamma + \alpha\gamma^2)\\ &+ 3\rho^2(\alpha\beta^2 + \beta\gamma^2 + \alpha^2\gamma) + 6\alpha\beta\gamma.\end{aligned} \tag{14.24}$$

We have

$$\begin{aligned}\sqrt{D} &= (\alpha - \beta)(\alpha - \gamma)(\beta - \gamma)\\ &= (\alpha^2\beta + \beta^2\gamma + \alpha\gamma^2) - (\alpha\beta^2 + \beta\gamma^2 + \alpha^2\gamma).\end{aligned}$$

Using this equation we see that (24) can be written

$$\alpha^3 + \beta^3 + \gamma^3 + 3\rho[\frac{1}{2}(S + \sqrt{D})] + 3\rho^2[\frac{1}{2}(S - \sqrt{D})] + 6\alpha\beta\gamma \tag{14.24$'$}$$

where for simplicity we have denoted by S the expression

$$(\alpha^2\beta + \beta^2\gamma + \alpha\gamma^2) + (\alpha\beta^2 + \beta\gamma^2 + \alpha^2\gamma).$$

Since S is symmetric in the roots, each of the expressions in (24$'$) is a symmetric polynomial in α, β and γ, hence is a polynomial in the elementary symmetric functions $s_1 = 0$, $s_2 = p$, and $s_3 = -q$. After a short calculation one finds

$$\alpha^3 + \beta^3 + \gamma^3 = -3q \qquad S = 3q$$

so that from (24$'$) we find ($\rho + \rho^2 = -1$ and $\rho - \rho^2 = \sqrt{-3}$)

$$\begin{aligned}\theta_1^3 &= -3q + \frac{3}{2}\rho(3q + \sqrt{D}) + \frac{3}{2}\rho^2(3q - \sqrt{D}) - 6q\\ &= \frac{-27}{2}q + \frac{3}{2}\sqrt{-3D}.\end{aligned} \tag{14.25}$$

Similarly, we find

$$\theta_2^3 = \frac{-27}{2}q - \frac{3}{2}\sqrt{-3D}. \tag{14.25$'$}$$

Equations (25) and (23) essentially give the solutions of our cubic. One small point remains, however, namely the issue of extracting the cube roots of the expressions in (25) to obtain θ_1 and θ_2. There are 3 possible cube roots, which might suggest a total of 9 expressions in (23). This is not the case since θ_1 and θ_2 are not independent (adjoining one of them already gives the Galois extension containing all of the roots). A computation like the one above (but easier) shows that

$$\theta_1\theta_2 = -3p \tag{14.26}$$

showing that the choice of cube root for θ_1 determines θ_2. Using $D = -4p^3 - 27q^2$, we obtain Cardano's explicit formulas, as follows.

Let

$$A = \sqrt[3]{\frac{-27}{2}q + \frac{3}{2}\sqrt{-3D}}$$

$$B = \sqrt[3]{\frac{-27}{2}q - \frac{3}{2}\sqrt{-3D}}$$

where the cube roots are chosen so that $AB = -3p$. Then the roots of the equation

$$y^3 + py + q = 0$$

are

$$\alpha = \frac{A+B}{3} \qquad \beta = \frac{\rho^2 A + \rho B}{3} \qquad \gamma = \frac{\rho A + \rho^2 B}{3} \tag{14.27}$$

where $\rho = -\frac{1}{2} + \frac{1}{2}\sqrt{-3}$.

Examples

(1) Consider the cubic equation $x^3 - x + 1 = 0$. The discriminant of this cubic is

$$D = -4(-1)^3 - 27(1)^2 = -23$$

which is not the square of a rational number, so the Galois group for this polynomial is S_3. Substituting into the formulas above we have

$$A = \sqrt[3]{\frac{-27}{2} + \frac{3}{2}\sqrt{69}}$$

$$B = \sqrt[3]{\frac{-27}{2} - \frac{3}{2}\sqrt{69}}$$

where we choose A to be the real cube root and then from $AB = 3$ we see that B is also real. The roots of the cubic are given by (27) and we see that there is one real root and two (conjugate) complex roots (which we could have determined without solving for the roots, of course).

(2) Consider the equation $x^3 + x^2 - 2x - 1 = 0$. Letting $x = s - 1/3$ the equation becomes $s^3 - \frac{7}{3}s - \frac{7}{27} = 0$. Multiplying through by 27 to clear denominators and letting $y = 3s$ we see that y satisfies the cubic equation

$$y^3 - 21y - 7 = 0.$$

The discriminant D for this cubic is

$$D = -4(-21)^3 - 27(-7)^2 = 3^6 7^2$$

which shows that the Galois group for this (Eisenstein at 7) cubic is A_3. Substituting into the formulas above we have

$$A = 3\sqrt[3]{\frac{7}{2} + \frac{21}{2}\sqrt{-3}}$$

$$B = 3\sqrt[3]{\frac{7}{2} - \frac{21}{2}\sqrt{-3}}$$

and the roots of our cubics can be expressed in terms of A and B using the formulas above. This cubic arises from trying to express a primitive 7^{th} root of unity ζ_7 in terms of radicals similar to the explicit formulas for the other roots of unity of small order (cf. the exercises).

In this case we have $g(-5) = -27$, $g(-1) = 13$, $g(0) = -7$ and $g(5) = 13$, so that this cubic has 3 *real* roots. The expressions above for these roots are sums of the conjugates of *complex* numbers. We shall see later that this is necessary, namely that it is impossible to solve for these real roots using only radicals involving real numbers.

A cubic with rational coefficients has either one real root and two complex conjugate imaginary roots or has three real roots. These two cases can be distinguished by the sign of the discriminant:

Suppose in the first case that the roots are a and $b \pm ic$ where a, b, and c are real and $c \neq 0$. Then

$$\begin{aligned}\sqrt{D} &= [a - (b + ic)][a - (b - ic)][(b + ic) - (b - ic)]\\ &= 2ic[(a - b)^2 + c^2]\end{aligned}$$

is purely imaginary, so that the discriminant D is negative. Then in the formulas for A and B above we may choose both to be real. The first root in (27) is then real and the second two are complex conjugates.

If all three roots are real, then clearly $\sqrt{D}$ is real, so $D \geq 0$ is a nonnegative real number. If $D = 0$ then the cubic has repeated roots. For $D > 0$ (sometimes called the *Casus irreducibilis*), the formulas for the roots involve radicals of nonreal numbers, as in Example 2. We now show that for irreducible cubics this is necessary. The exercises outline the proof of the following generalization: if all the roots of the irreducible polynomial $f(x) \in \mathbb{Q}[x]$ are real and if one of these roots can be expressed by *real* radicals, then the degree of $f(x)$ is a power of 2, the Galois group of $f(x)$ is a 2-group, and the roots of $f(x)$ can be constructed by straightedge and compass.

Suppose that the irreducible cubic $f(x)$ has three real roots and that it were possible to express one of these roots by radicals involving only real numbers. Then the splitting field for the cubic would be contained in a root extension

$$\mathbb{Q} = K_0 \subset K_1 = \mathbb{Q}(\sqrt{D}) \subset \cdots \subset K_i \subset K_{i+1} \subset \cdots \subset K_s = K$$

where each field K_i, $i = 0, 1, \ldots, s$, is contained in the real numbers $\mathbb{R}$ and $s \geq 2$ since the quadratic extension $\mathbb{Q}(\sqrt{D})$ cannot contain the root of an irreducible cubic. We have begun this root extension with $\mathbb{Q}(\sqrt{D})$ because over this field the Galois group of the polynomial is cyclic of degree 3.

Note that for any field F the extension $F(\sqrt[mn]{a})$ of F can be obtained by two smaller simple radical extensions: let

$$F_1 = F(\sqrt[n]{a})$$

and let $b = \sqrt[n]{a} \in F_1$, so that

$$F(\sqrt[mn]{a}) = F_1(\sqrt[m]{b}).$$

We may therefore always assume our radical extensions are of the form $F(\sqrt[p]{a})$ where p is a prime.

Suppose now that F is a subfield of the real numbers $\mathbb{R}$ and let a be an element of F. Let p be a prime and let $\alpha = \sqrt[p]{a}$ denote a real p^{th} root of a. Then $[F(\sqrt[p]{a}) : F]$ must be either 1 or p, as follows. The conjugates of α over F all differ from α by a p^{th} root of unity. It follows that the constant term of the minimal polynomial of α over F is $\alpha^d \zeta$ where $d = [F(\sqrt[p]{a}) : F]$ is the degree of the minimal polynomial and ζ is some p^{th} root of unity. Since α is real and $\alpha^d \zeta \in F$ is real, it follows that $\zeta = \pm 1$, so that $\alpha^d \in F$. Then, if $d \neq p$, $\alpha^d \in F$ and $\alpha^p = a \in F$ implies $\alpha \in F$, so $d = 1$.

Hence we may assume for the radical extensions above that $[K_{i+1} : K_i]$ is a prime p_i and $K_{i+1} = K_i(\sqrt[p_i]{a_i})$ for some $a_i \in K_i$, $i = 0, 1, \dots, s-1$. In other words, the original tower of real radical extensions can be refined to a tower where each of the successive radical extensions has prime degree.

If any field containing $\sqrt{D}$ contains one of the roots of $f(x)$ then it contains the splitting field for $f(x)$, hence contains all the roots of the cubic. We suppose s is chosen so that K_{s-1} does not contain any of the roots of the cubic.

Consider the extension K_s/K_{s-1}. The field K_s contains all the roots of the cubic $f(x)$ and the field K_{s-1} contains none of these roots. It follows that $f(x)$ is irreducible over K_{s-1}, so $[K_s : K_{s-1}]$ is divisible by 3. Since we have reduced to the case where this extension degree is a prime, it follows that the extension degree is precisely 3 and that the extension K_s/K_{s-1} is Galois (being the splitting field of $f(x)$ over K_{s-1}). Since also $K_s = K_{s-1}(\sqrt[3]{a})$ for some $a \in K_{s-1}$, the Galois extension K_s must also contain the other cube roots of a. This implies that K_s contains ρ, a primitive 3^{rd} root of unity. This contradicts the assumption that K_s is a subfield of $\mathbb{R}$ and shows that it is impossible to express the roots of this cubic in terms of real radicals only.

Solution of Quartic Equations by Radicals

Consider now the case of a quartic polynomial

$$f(x) = x^4 + ax^3 + bx^2 + cx + d$$

which under the substitution $x = y - a/4$ becomes the quartic

$$g(y) = y^4 + py^2 + qy + r$$

with

$$\begin{aligned} p &= \frac{1}{8}(-3a^2 + 8b) \\ q &= \frac{1}{8}(a^3 - 4ab + 8c) \\ r &= \frac{1}{256}(-3a^4 + 16a^2b - 64ac + 256d). \end{aligned}$$

Let the roots of $g(y)$ be α_1, α_2, α_3, and α_4. The resolvent cubic introduced in the previous section for this quartic is

$$h(x) = x^3 - 2px^2 + (p^2 - 4r)x + q^2$$

and has roots

$$\begin{aligned}\theta_1 &= (\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)\\ \theta_2 &= (\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4)\\ \theta_3 &= (\alpha_1 + \alpha_4)(\alpha_2 + \alpha_3).\end{aligned}$$

The Galois group of the splitting field for $f(x)$ (or $g(y)$) over the splitting field of the resolvent cubic $h(x)$ is the Klein 4-group. Such extensions are biquadratic, which means that it is possible to solve for the roots $\alpha_1, \alpha_2, \alpha_3$, and α_4 in terms of square roots of expressions involving the roots θ_1, θ_2, and θ_3 of the resolvent cubic. In this case we evidently have

$$(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4) = \theta_1 \qquad (\alpha_1 + \alpha_2) + (\alpha_3 + \alpha_4) = 0$$

which gives

$$\alpha_1 + \alpha_2 = \sqrt{-\theta_1} \qquad \alpha_3 + \alpha_4 = -\sqrt{-\theta_1}.$$

Similarly,

$$\begin{aligned}\alpha_1 + \alpha_3 &= \sqrt{-\theta_2} &\qquad \alpha_2 + \alpha_4 &= -\sqrt{-\theta_2}\\ \alpha_1 + \alpha_4 &= \sqrt{-\theta_3} &\qquad \alpha_2 + \alpha_3 &= -\sqrt{-\theta_3}.\end{aligned}$$

An easy computation shows that $\sqrt{-\theta_1}\sqrt{-\theta_2}\sqrt{-\theta_3} = -q$, so that the choice of two of the square roots determines the third. Since $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$, if we add the left-hand equations above we obtain $2\alpha_1$, and similarly we may solve for the other roots of $g(y)$. We find

$$\begin{aligned}2\alpha_1 &= \sqrt{-\theta_1} + \sqrt{-\theta_2} + \sqrt{-\theta_3}\\ 2\alpha_2 &= \sqrt{-\theta_1} - \sqrt{-\theta_2} - \sqrt{-\theta_3}\\ 2\alpha_3 &= -\sqrt{-\theta_1} + \sqrt{-\theta_2} - \sqrt{-\theta_3}\\ 2\alpha_4 &= -\sqrt{-\theta_1} - \sqrt{-\theta_2} + \sqrt{-\theta_3}\end{aligned}$$

which reduces the solution of the quartic equation to the solution of the associated resolvent cubic.

EXERCISES

1. Use Cardano's Formulas to solve the equation $x^3 + x^2 - 2 = 0$. In particular show that the equation has the real root

$$\frac{1}{3}(\sqrt[3]{26 + 15\sqrt{3}} + \sqrt[3]{26 - 15\sqrt{3}} - 1).$$

Show directly that the roots of this cubic are $1, -1 \pm i$. Explain this by proving that

$$\sqrt[3]{26 + 15\sqrt{3}} = 2 + \sqrt{3} \quad \sqrt[3]{26 - 15\sqrt{3}} = 2 - \sqrt{3}$$

so that

$$\sqrt[3]{26+15\sqrt{3}}+\sqrt[3]{26-15\sqrt{3}}=4.$$

2. Let ζ_7 be a primitive 7th root of unity and let $\alpha = \zeta + \zeta^{-1}$.

(a) Show that ζ_7 is a root of the quadratic $z^2 - \alpha z + 1$ over $\mathbb{Q}(\alpha)$.

(b) Show using the minimal polynomial for ζ_7 that α is a root of the cubic x^3+x^2-2x-1.

(c) Use (a) and (b) together with the explicit solution of the cubic in (b) in the text to express ζ_7 in terms of radicals similar to the expressions given earlier for the other roots of unity of small order. (The complicated nature of the expression explains why we did not include ζ_7 earlier in our list of explicit roots of unity.)

3. Let F be a field of characteristic $\neq 2$. State and prove a necessary and sufficient condition on $\alpha, \beta \in F$ so that $F(\sqrt{\alpha}) = F(\sqrt{\beta})$. Use this to determine whether $\mathbb{Q}(\sqrt{1-\sqrt{2}}) = \mathbb{Q}(i, \sqrt{2})$.

4. Let $K = \mathbb{Q}(\sqrt[n]{a})$, where $a \in \mathbb{Q}, a > 0$ and suppose $[K : \mathbb{Q}] = n$ (i.e., $x^n - a$ is irreducible). Let E be any subfield of K and let $[E : \mathbb{Q}] = d$. Prove that $E = \mathbb{Q}(\sqrt[d]{a})$. [Consider $N_{K/E}(\sqrt[n]{a}) \in E$.]

5. Let K be as in the previous exercise. Prove that if n is odd then K has no nontrivial subfields which are Galois over $\mathbb{Q}$ and if n is even then the only nontrivial subfield of K which is Galois over $\mathbb{Q}$ is $\mathbb{Q}(\sqrt{a})$.

6. Let L be the Galois closure of K in the previous two exercises (i.e., the splitting field of $x^n - a$). Prove that $[L : \mathbb{Q}] = n\varphi(n)$ or $\frac{1}{2}n\varphi(n)$. [Note that $\mathbb{Q}(\zeta_n) \cap K$ is a Galois extension of $\mathbb{Q}$.]

7. (*Kummer Generators for Cyclic Extensions*) Let F be a field of characteristic not dividing n containing the nth roots of unity and let K be a cyclic extension of degree d dividing n. Then $K = F(\sqrt[n]{a})$ for some nonzero $a \in F$. Let σ be a generator for the cyclic group $\mathrm{Gal}(K/F)$.

(a) Show that $\sigma(\sqrt[n]{a}) = \zeta\sqrt[n]{a}$ for some primitive dth root of unity ζ.

(b) Suppose $K = F(\sqrt[n]{a}) = F(\sqrt[n]{b})$. Use (a) to show that $\dfrac{\sigma(\sqrt[n]{a})}{\sqrt[n]{a}} = \left(\dfrac{\sigma(\sqrt[n]{b})}{\sqrt[n]{b}}\right)^i$ for some integer i relatively prime to d. Conclude that σ fixes the element $\dfrac{\sqrt[n]{a}}{(\sqrt[n]{b})^i}$ so this is an element of F.

(c) Prove that $K = F(\sqrt[n]{a}) = F(\sqrt[n]{b})$ if and only if $a = b^i c^n$ and $b = a^j d^n$ for some $c, d \in F$, i.e., if and only if a and b generate the same subgroup of $F^\times$ modulo nth powers.

8. Let p, q and r be primes in $\mathbb{Z}$ with $q \neq r$. Let $\sqrt[p]{q}$ denote any root of $x^p - q$ and let $\sqrt[p]{r}$ denote any root of $x^p - r$. Prove that $\mathbb{Q}(\sqrt[p]{q}) \neq \mathbb{Q}(\sqrt[p]{r})$.

9. (*Artin–Schreier Extensions*) Let F be a field of characteristic p and let K be a cyclic extension of F of degree p. Prove that $K = F(\alpha)$ where α is a root of the polynomial $x^p - x - a$ for some $a \in F$. [Note that $\mathrm{Tr}_{K/F}(-1) = 0$ since F is of characteristic p so that $-1 = \alpha - \sigma\alpha$ for some $\alpha \in K$ where σ is a generator of $\mathrm{Gal}(K/F)$ by Exercise 26 of Section 2. Show that $a = \alpha^p - \alpha$ is an element of F.] Note that since F contains the pth roots of unity (namely, 1) that this completes the description of all cyclic extensions of prime degree p over fields containing the pth roots of unity in all characteristics.

10. Let $K = \mathbb{Q}(\zeta_p)$ be the cyclotomic field of pth roots of unity for the prime p and let

$G = \mathrm{Gal}(K/\mathbb{Q})$. Let ζ denote any p^{th} root of unity. Prove that $\sum_{\sigma\in G}\sigma(\zeta)$ (the trace from K to $\mathbb{Q}$ of ζ) is -1 or $p-1$ depending on whether ζ is or is not a primitive p^{th} root of unity.

11. (*The Classical Gauss Sum*) Let $K = \mathbb{Q}(\zeta_p)$ be the cyclotomic field of p^{th} roots of unity for the odd prime p, viewed as a subfield of $\mathbb{C}$, and let $G = \mathrm{Gal}(K/\mathbb{Q})$. Let H denote the subgroup of index 2 in the cyclic group G. Define $\eta_0 = \sum_{\tau\in H}\tau(\zeta_p)$, $\eta_1 = \sum_{\tau\in\sigma H}\tau(\zeta_p)$, where σ is a generator of $\mathrm{Gal}(K/\mathbb{Q})$ (the two *periods* of ζ_p with respect to H, i.e., the sum of the conjugates of ζ_p with respect to the two cosets of H in G, cf. Section 5).

(a) Prove that $\sigma(\eta_0) = \eta_1$, $\sigma(\eta_1) = \eta_0$ and that

$$\eta_0 = \sum_{a=\text{square}} \zeta_p^a \quad , \quad \eta_1 = \sum_{b\neq\text{square}} \zeta_p^b.$$

where the sums are over the squares and nonsquares (respectively) in $(\mathbb{Z}/p\mathbb{Z})^\times$. [Observe that H is the subgroup of squares in $(\mathbb{Z}/p\mathbb{Z})^\times$.]

(b) Prove that $\eta_0+\eta_1 = (\zeta_p, 1) = -1$ and $\eta_0-\eta_1 = (\zeta_p, -1)$ where $(\zeta_p, 1)$ and $(\zeta_p, -1)$ are two of the Lagrange resolvents of ζ_p.

(c) Let $g = \sum_{i=0}^{p-1}\zeta_p^{i^2}$ (the classical *Gauss sum*). Prove that

$$g = (\zeta_p, -1) = \sum_{i=0}^{p-2}(-1)^i\,\sigma^i(\zeta_p).$$

(d) Prove that $\tau g = g$ if $\tau \in H$ and $\tau g = -g$ if $\tau \notin H$. Conclude in particular that $[\mathbb{Q}(g) : \mathbb{Q}] = 2$. Recall that complex conjugation is the automorphism σ_{-1} on K (cf. Exercise 7 of Section 5). Conclude that $\bar{g} = g$ if -1 is a square mod p (i.e., if $p \equiv 1 \bmod 4$) and $\bar{g} = -g$ if -1 is not a square mod p (i.e., if $p \equiv 3 \bmod 4$) where $\bar{g}$ denotes the complex conjugate of g.

(e) Prove that $g\bar{g} = p$. [The complex conjugate of a root of unity is its reciprocal. Then $\bar{g} = \sum_{j=0}^{p-2}(-1)^j\,(\sigma^j(\zeta_p))^{-1}$ gives

$$\begin{aligned} g\bar{g} &= \sum_{i,j=0}^{p-2}(-1)^i(-1)^j\frac{\sigma^i(\zeta_p)}{\sigma^j(\zeta_p)} = \sum_{i,j=0}^{p-2}(-1)^{i-j}\sigma^j\left[\frac{\sigma^{i-j}(\zeta_p)}{\zeta_p}\right] \\ &= \sum_{k=0}^{p-2}(-1)^k\sum_{j=0}^{p-2}\sigma^j\left[\frac{\sigma^k(\zeta_p)}{\zeta_p}\right] \end{aligned}$$

where $k = i-j$. If $k=0$ the element $\dfrac{\sigma^k(\zeta_p)}{\zeta_p}$ is 1, and if $k\neq 0$ then this is a primitive p^{th} root of unity. Use the previous exercise to conclude that the inner sum is $p-1$ when $k=0$ and is -1 otherwise.]

(f) Conclude that $g^2 = (-1)^{(p-1)/2}p$ and that $\mathbb{Q}(\sqrt{(-1)^{(p-1)/2}p}\,)$ is the unique quadratic subfield of $\mathbb{Q}(\zeta_p)$. (Cf. also Exercise 33 of Section 6.)

12. Let L be the Galois closure of the finite extension $\mathbb{Q}(\alpha)$ of $\mathbb{Q}$. For any prime p dividing the order of $\mathrm{Gal}(L/\mathbb{Q})$ prove there is a subfield F of L with $[L : F] = p$ and $L = F(\alpha)$.

13. Let F be a subfield of the real numbers $\mathbb{R}$. Let a be an element of F and let $K = F(\sqrt[n]{a})$ where $\sqrt[n]{a}$ denotes a real n^{th} root of a. Prove that if L is any Galois extension of F contained in K then $[L : F] \leq 2$.

14. This exercise shows that in general it is necessary to use complex numbers when expressing real roots in terms of radicals and generalizes the *Casus irreducibilis* of cubic equations.

Let $f(x) \in \mathbb{Q}[x]$ be an irreducible polynomial all of whose roots are real. Suppose further that one of the roots, α, of $f(x)$ can be expressed in terms of *real* radicals (i.e., there is a root extension of real fields $\mathbb{Q} = K_0 \subset K_1 \subset \dots \subset K_m \subset \mathbb{R}$ with $K_{i+1} = K_i(\sqrt[n_i]{a_i})$, $i = 1, 2, \dots, m-1$, for some integers n_i and some $a_i \in K_i$ and $\alpha \in K_m$). Prove that the Galois group of $f(x)$ is a 2-group. Conclude in particular that the degree of $f(x)$ is a power of 2 and that the real roots of such a polynomial can be expressed entirely in terms of real radicals if and only if these roots can be constructed by straightedge and compass. [The argument is similar to the case of cubics. Let $L \in \mathbb{R}$ be the Galois closure of $\mathbb{Q}(\alpha)$ and suppose the order of $\mathrm{Gal}(L/\mathbb{Q})$ is divisible by some odd prime p. Let F be a subfield of L with $[L : F] = p$ and $L = F(\alpha)$ (by Exercise 12) and consider the composite fields $K_i' = FK_i$, $i = 0, 1, \dots, m$. These are again real radical extensions and by the argument in the text for the *Casus irreducibilis*, we may assume each $[K_{i+1}' : K_i']$ is a prime. Since $\alpha \notin F = FK_0$, there is an integer s with $\alpha \notin K_{s-1}'$, $\alpha \in K_s'$. Since the extensions are of prime degree, we have $K_s' = K_{s-1}'(\alpha)$. Since $L = F(\alpha)$ is Galois of degree p, K_s' is a Galois extension of K_{s-1}' of degree p, contradicting the previous exercise.]

15. (*'Cardano's Formulas' for a Cubic in Characteristic 2)* Suppose $f(x) = x^3 + px + q$ is an irreducible cubic over a field of characteristic 2. Let ρ be a primitive 3rd root of unity and let θ, θ' be the roots of the quadratic $x^2 + qx + (p^3 + q^2)$ (cf. Exercise 50 of Section 6). Let θ_1 and θ_2 be cube roots of $\rho q + \theta$ and $\rho q + \theta'$, respectively, where the cube roots are chosen so that $\theta_1\theta_2 = p$. Prove that the roots of $f(x)$ are given by $\alpha = \theta_1 + \theta_2$, $\beta = \rho\alpha + \theta_1$, and $\gamma = \rho\alpha + \theta_2 = \alpha + \beta$.

16. Let a be a nonzero rational number.

(a) Determine when the extension $\mathbb{Q}(\sqrt{ai})$ $(i^2 = -1)$ is of degree 4 over $\mathbb{Q}$.

(b) When $K = \mathbb{Q}(\sqrt{ai})$ is of degree 4 over $\mathbb{Q}$ show that K is Galois over $\mathbb{Q}$ with the Klein 4-group as Galois group. In this case determine the quadratic extensions of $\mathbb{Q}$ contained in K.

17. Let $D \in \mathbb{Z}$ be a squarefree integer and let $a \in \mathbb{Q}$ be a nonzero rational number. Show that $\mathbb{Q}(\sqrt{a\sqrt{D}})$ cannot be a cyclic extension of degree 4 over $\mathbb{Q}$.

18. Let $D \in \mathbb{Z}$ be a squarefree integer and let $a \in \mathbb{Q}$ be a nonzero rational number. Prove that if $\mathbb{Q}(\sqrt{a\sqrt{D}})$ is Galois over $\mathbb{Q}$ then $D = -1$.

19. Let $D \in \mathbb{Z}$ be a squarefree integer and let $K = \mathbb{Q}(\sqrt{D})$.

(a) Prove that if $D = s^2 + t^2$ is the sum of two rational squares then there exists an extension $L/\mathbb{Q}$ containing K which is Galois over $\mathbb{Q}$ with a cyclic Galois group of order 4. [Consider the extension $\mathbb{Q}(\sqrt{D + s\sqrt{D}})$.] (Note also that D is the sum of two rational squares if and only if D is also the sum of two integer squares, so one may assume s and t are integral without loss.)

(b) Prove conversely that if K can be embedded in a cyclic extension L of degree 4 as in (a) then D is the sum of two squares. [One approach: (i) observe first that L is quadratic over K, so $L = K(\sqrt{a + b\sqrt{D}})$ for some $a, b \in \mathbb{Q}$, (ii) show that L contains the quadratic subfield $\mathbb{Q}(\sqrt{a^2 - b^2 D})$, which must be $\mathbb{Q}(\sqrt{D})$ if $L/\mathbb{Q}$ is cyclic, and use Exercise 7.]

(c) Conclude in particular that $\mathbb{Q}(\sqrt{3})$ is not a subfield of any cyclic extension of degree 4 over $\mathbb{Q}$. Similarly conclude that the fields $\mathbb{Q}(\sqrt{D})$ for squarefree integers $D < 0$ are never contained in cyclic extensions of degree 4 over $\mathbb{Q}$ (this gives an alternate proof for Exercise 19, Section 6).

20. Let p be a prime. Show that any solvable subgroup of S_p of order divisible by p is contained in the normalizer of a Sylow p-subgroup of S_p (a Frobenius group of order $p(p-1)$). Conclude that an irreducible polynomial $f(x) \in \mathbb{Q}[x]$ of degree p is solvable by radicals if and only if its Galois group is contained in the Frobenius group of order $p(p-1)$. [Let $G \leq S_p$ be a solvable subgroup of order divisible by p. Then G contains a p-cycle, hence is transitive on $\{1, 2, \ldots, p\}$. Let $H < G$ be the stabilizer in G of the element 1, so H has index p in G. Show that H contains no nontrivial normal subgroups of G (note that the conjugates of H are the stabilizers of the other points). Let $G^{(n-1)}$ be the last nontrivial subgroup in the derived series for G. Show that $H \cap G^{(n-1)} = 1$ and conclude that $|G^{(n-1)}| = p$, so that the Sylow p-subgroup of G (which is also a Sylow p-subgroup in S_p) is normal in G.]

21. (*Criterion for the Solvability of a Quintic*) By the previous exercise, an irreducible polynomial $f(x)$ in $\mathbb{Q}[x]$ of degree 5 can be solved by radicals if and only if its Galois group (considered as a subgroup of S_5) is contained in the Frobenius group of order 20. It is known that this is the case if and only if an associated polynomial $g(x)$ of degree 6 has a rational root (cf. Dummit, *Solving Solvable Quintics*, Math. Comp., 57(1991), pp. 387–401). If the quintic is in the general form (where a translation is performed so that the coefficient of x^4 is zero)

$$f(x) = x^5 + px^3 + qx^2 + rx + s \qquad p, q, r, s \in \mathbb{Q}$$

then the associated polynomial of degree 6 is

$$\begin{aligned} g(x) = x^6 &+ 8rx^5 + (2pq^2 - 6p^2r + 40r^2 - 50qs)\,x^4 \\ &+ (-2q^4 + 21pq^2r - 40p^2r^2 + 160r^3 - 15p^2qs - 400qrs + 125ps^2)\,x^3 \\ &+ (p^2q^4 - 8q^4r + 9p^4r^2 - 136p^2r^3 + 625q^2s^2 + 400r^4 - 6p^3q^2r \\ &\quad + 76pq^2r^2 - 50pq^3s - 1400qr^2s + 500prs^2 + 90p^2qrs)\,x^2 \\ &+ (-108p^5s^2 + 32p^4r^3 - 256p^2r^4 - 3125s^4 + 512r^5 - 2pq^6 + 3q^4r^2 \\ &\quad - 58q^5s + 2750q^2rs^2 - 31p^3q^3s - 500pr^2s^2 + 19p^2q^4r \\ &\quad - 51p^3q^2r^2 + 76pq^2r^3 - 2400qr^3s - 325p^2q^2s^2 + 525p^3rs^2 \\ &\quad + 625pqs^3 + 117p^4qrs + 105pq^3rs + 260p^2qr^2s)\,x \\ &+ (q^8 + 256r^6 + 17q^4r^3 - 27p^7s^2 - 4p^6r^3 + 48p^4r^4 - 192p^2r^5 \\ &\quad + 3125p^2s^4 - 9375rs^4 - 1600qr^4s - 99p^5rs^2 - 125pq^4s^2 \\ &\quad - 124q^5rs + 3250q^2r^2s^2 - 2000pr^3s^2 - 13pq^6r + p^5q^2r^2 \\ &\quad + 65p^2q^4r^2 - 128p^3q^2r^3 - 16pq^2r^4 - 4p^5q^3s - 12p^2q^5s \\ &\quad - 150p^4q^2s^2 + 1200p^3r^2s^2 + 18p^6qrs + 12p^3q^3rs + 196p^4qr^2s \\ &\quad + 590pq^3r^2s - 160p^2qr^3s - 725p^2q^2rs^2 - 1250pqrs^3). \end{aligned}$$

In the particular case where $f(x) = x^5 + Ax + B$ this polynomial is simply

$$g(x) = x^6 + 8Ax^5 + 40A^2x^4 + 160A^3x^3 + 400A^4x^2 + (512A^5 - 3125B^4)x - 9375AB^4 + 256A^6.$$

(a) Use this criterion to prove that the Galois group over $\mathbb{Q}$ of the polynomial $x^5 - 5x + 12$ is the dihedral group of order 10. [Show the associated sixth degree polynomial is

$$x^6 - 40x^5 + 1000x^4 - 20000x^3 + 250000x^2 - 66400000x + 976000000$$

and has $x = 40$ as a rational root. Cf. also Exercise 35 in Section 6.]

(b) Use this criterion to prove that $x^5 - x - 1$ is not solvable by radicals.

14.8 COMPUTATION OF GALOIS GROUPS OVER $\mathbb{Q}$

In the determination of the Galois groups of polynomials of degrees ≤ 4 in Section 6 and in the determination of the Galois group of the polynomial $x^5 - 6x + 3$ in the previous section we observed that it was possible to obtain useful information regarding the Galois group from the *cycle types* of the automorphisms as elements in S_n. This is very useful in computing Galois groups of polynomials over $\mathbb{Q}$ and we now briefly describe the theoretical justification.

Let $f(x)$ be a polynomial with rational coefficients. In determining the Galois group of $f(x)$ we may assume that $f(x)$ is separable and has integer coefficients. Then the discriminant D of $f(x)$ is an integer and is nonzero.

For any prime p, consider the reduction $\overline{f}(x) \in \mathbb{F}_p[x]$ of $f(x)$ modulo p. If p divides D then the reduced polynomial $\overline{f}(x)$ has discriminant $\overline{D} = 0$ in $\mathbb{F}_p$, so is not separable.

If p does not divide D, then $\overline{f}(x)$ is a separable polynomial over $\mathbb{F}_p$ and we can factor $\overline{f}(x)$ into distinct irreducibles

$$\overline{f}(x) = \overline{f}_1(x)\overline{f}_2(x)\cdots\overline{f}_k(x) \qquad \text{in } \mathbb{F}_p[x].$$

Let n_i be the degree of $\overline{f}_i(x)$, $i = 1, 2, \ldots, k$.

The importance of this reduction is provided by the following theorem from algebraic number theory which is an elementary consequence of the study of the arithmetic in finite extensions of $\mathbb{Q}$ (and which we take for granted).

Theorem. For any prime p not dividing the discriminant D of $f(x) \in \mathbb{Z}[x]$, the Galois group over $\mathbb{F}_p$ of the reduction $\overline{f}(x) = f(x) \pmod{p}$ is permutation group isomorphic to a subgroup of the Galois group over $\mathbb{Q}$ of $f(x)$.

The meaning of the statement "permutation group isomorphic" in the theorem is that not only is the Galois group of the reduction $\overline{f}(x)$ mod p of $f(x)$ isomorphic to a subgroup of the Galois group of $f(x)$ but that there is an ordering of the roots of $\overline{f}(x)$ and of $f(x)$ (depending on p) so that under this isomorphism the action of the corresponding automorphisms as permutations of these roots is the same. In particular there are automorphisms in the Galois group of $f(x)$ with the same cycle types as the automorphisms of $\overline{f}(x)$.

The Galois group of $\overline{f}(x)$ is a *cyclic* group since every finite extension of $\mathbb{F}_p$ is a cyclic extension. Let σ be a generator for this Galois group over $\mathbb{F}_p$ (for example, the Frobenius automorphism). The roots of $\overline{f}_1(x)$ are permuted amongst themselves by the Galois group, and given any two of these roots there is a Galois automorphism taking the first root to the second (recall that the group is said to be *transitive* on the roots when this is the case). Similarly, the Galois group permutes the roots of each of the factors $\overline{f}_i(x)$, $i = 1, 2, \ldots, k$ transitively. Since these factors are relatively prime we also see that no root of one factor is mapped to a root of any other factor by any element of the Galois group.

View σ as an element in S_n by labelling the n roots of $\overline{f}(x)$ and consider the cycle decomposition of σ, which is a product of k distinct permutations since σ permutes the roots of each of the factors $\overline{f}_i(x)$ amongst themselves. By the observations we just made, the action of σ on the roots of $\overline{f}_1(x)$ must be a cycle of length n_i since otherwise the powers of σ could not be transitive on the roots of $\overline{f}_1(x)$. Similarly the action of σ on the roots of $\overline{f}_i(x)$ gives a cycle of length n_i, $i = 1, 2, \ldots, k$.

We see that the automorphism σ generating the Galois group of $\overline{f}(x)$ has cycle decomposition $(n_1, n_2, \ldots, n_k)$ where $n_1, n_2, \ldots, n_k$ are the degrees of the irreducible factors of $f(x)$ reduced modulo p, which gives us the following result.

Corollary 41. For any prime p not dividing the discriminant of $f(x) \in \mathbb{Z}[x]$, the Galois group of $f(x)$ over $\mathbb{Q}$ contains an element with cycle decomposition $(n_1, n_2, \ldots, n_k)$ where $n_1, n_2, \ldots, n_k$ are the degrees of the irreducible factors of $f(x)$ reduced modulo p.

Example

Consider the polynomial $x^5 - x - 1$. The discriminant of this polynomial is $2869 = 19 \cdot 151$ so we reduce at primes $\neq 19, 151$. Reducing mod 2 the polynomial $x^5 - x - 1$ factors as $(x^2 + x + 1)(x^3 + x^2 + 1)$ (mod 2) so the Galois group has a (2,3)-cycle. Cubing this element we see the Galois group contains a transposition.

Reducing mod 3 the polynomial is irreducible, as follows: $x^5 - x - 1$ has no roots mod 3 so if it were reducible mod 3 then it would have an irreducible quadratic factor, hence would have a factor in common with $x^9 - x$ (which is the product of all irreducible polynomials of degrees 1 and 2 over $\mathbb{F}_3$), hence a factor in common with either $x^4 - 1$ or $x^4 + 1$, hence a factor in common with either $x^5 - x$ or $x^5 + x$, hence a factor in common with either -1 or $2x + 1$ which it obviously does not. This shows both that $x^5 - x - 1$ is irreducible in $\mathbb{Z}[x]$ and that there is a 5-cycle in its Galois group.

Since S_5 is generated by any 5-cycle and any transposition, it follows that the Galois group of $x^5 - x - 1$ is S_5 (so in particular this polynomial cannot be solved by radicals, (cf. Exercise 21 of Section 7).

The arguments in the example above indicate how to construct polynomials with S_n as Galois group. We use the fact that a transitive subgroup of S_n containing a transposition and an $n - 1$-cycle is S_n. Let f_1 be an irreducible polynomial of degree n over $\mathbb{F}_2$. Let $f_2 \in \mathbb{F}_3[x]$ be the product of an irreducible polynomial of degree 2 with irreducible polynomials of odd degree (for example, an irreducible polynomial of degree $n - 3$ and x if n is even and an irreducible polynomial of degree $n - 2$ if n is odd). Let $f_3 \in \mathbb{F}_5[x]$ be the product of x with an irreducible polynomial of degree $n - 1$. Finally, let $f(x) \in \mathbb{Z}[x]$ be any polynomial with

$$\begin{aligned} f(x) &\equiv f_1(x) \pmod 2 \\ &\equiv f_2(x) \pmod 3 \\ &\equiv f_3(x) \pmod 5. \end{aligned}$$

The reduction of $f(x)$ mod 2 shows that $f(x)$ is irreducible in $\mathbb{Z}[x]$, hence the Galois group is transitive on the n roots of $f(x)$. Raising the element given by the factorization of $f(x)$ mod 3 to a suitable odd power shows the Galois group contains a transposition. The factorization mod 5 shows the Galois group contains an $n - 1$-cycle, hence the Galois group is S_n.

Proposition 42. For each $n \in \mathbb{Z}^+$ there exist infinitely many polynomials $f(x) \in \mathbb{Z}[x]$ with S_n as Galois group over $\mathbb{Q}$.

There are extremely efficient algorithms for factoring polynomials $f(x) \in \mathbb{Z}[x]$ modulo p (cf. Exercises 12 to 17 of Section 3), so the corollary above is an effective procedure for determining some of the cycle types of the elements of the Galois group. In using Corollary 41 some care should be taken not to assume that a *particular* cycle is an element of the Galois group. For example, one factorization might imply the existence of a (2,2) cycle, say (12)(34) and another factorization imply the existence of a transposition. One cannot conclude that the transposition is necessarily (12), however (nor (34), nor (13), etc.). The choice of (12)(34) to represent the first cycle fixes a particular ordering on the roots and this may not be the ordering with respect to which the transposition appears as (12).

Corollary 41 is particularly efficient in determining when the Galois group is large (e.g., S_n), since a transitive group containing sufficiently many cycle types must be S_n (for example, a transitive subgroup of S_n containing a transposition and an $n-1$-cycle is S_n, as used above). The most difficult Galois groups to determine in this way are the *small* Galois groups (e.g., a cyclic group of order n), since factorization after factorization will produce only elements of orders dividing n and one is not sure whether there will be some p yet to come producing a cycle type inconsistent with the assumption of a cyclic Galois group. If one could "compute forever" one could at least be sure of the precise distribution of cycle types among the elements of the Galois group in the following sense: suppose the Galois group $G \subseteq S_n$ has order N and that there are n_T elements of G with cycle type T (e.g., (2,2)-cycles, transpositions, etc.) so that the "density" of cycle type T in G is $d_T = n_T/N$. Then it is possible to define a density on the set of prime numbers (so that it makes sense to speak of "1/2" the primes, etc.) and we have the following result (which relies on the Tchebotarov Density Theorem in algebraic number theory).

Theorem. The density of primes p for which $f(x)$ splits into type T modulo p is precisely d_T.

This says that if we knew the factorization of $f(x)$ modulo every prime we could at least determine the number of elements of G with a given cycle type. Unfortunately, even this would not be sufficient to determine G (up to isomorphism): it is known that there are nonisomorphic groups containing the same number of elements of all cycle types (there are two nonisomorphic groups of order 96 in S_8 both having cycle type distributions: 1 1-cycle, 6 (2,2)-cycles, 13 (2,2,2,2)-cycles, 32 (3,3)-cycles, 12 (4,4)-cycles, 32 (2,6)-cycles). There are infinitely many such examples (the regular representation of the elementary abelian group of order p^3 and for the nonabelian group of order p^3 of exponent p give two nonisomorphic groups in S_{p^3} whose nonidentity elements are all the product of p^2 p-cycles for any prime p).

In practice one uses the factorizations of $f(x)$ modulo small primes to get an idea of the probable Galois group (based on the previous result). One then tries to prove this is indeed the Galois group — often a difficult problem. For polynomials of small degree, definitive algorithms exist, based in part on the computation of *resolvent* polynomials.

These are analogues of the cubic resolvent used in the previous sections to determine the Galois group of quartic polynomials. These resolvent polynomials have rational coefficients and have as roots certain combinations of the roots of $f(x)$ (similar to the combinations $(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)$ for the cubic resolvent). One then determines the factorization of these resolvent polynomials to obtain information on the Galois group of $f(x)$ — for example the existence of a linear factor implies the Galois group lies in the stabilizer in S_n of the combination of the roots of $f(x)$ chosen (for example, the dihedral group of order 8 for our resolvent cubic). It should be observed, however, that the degree of the resolvent polynomials constructed, unlike the situation of the resolvent cubic for quartic polynomials, are in general much larger than the degree of $f(x)$. The effectiveness of this computational technique also depends heavily on the explicit knowledge of the possible transitive subgroups of S_n. For $n = 2, 3, \dots, 8$ the number of isomorphism classes of transitive subgroups of S_n is 1, 2, 5, 5, 16, 7, 50, respectively. There is a great deal of interest in the computation of Galois groups, motivated in part by the problem of determining which groups occur as Galois groups over $\mathbb{Q}$.

We illustrate these techniques with some easier examples (from *The Computation of Galois Groups*, L. Soicher, Master's Thesis, Concordia University, Montreal, 1981).

Examples

(1) There are 5 isomorphism classes of transitive subgroups of S_5 given by the groups Z_5, D_{10}, F_{20}, the so-called Frobenius group of order 20 (the Galois group of $x^5 - 2$ with generators $(1\,2\,3\,4\,5)$ and $(2\,3\,5\,4)$ in S_5), A_5 and S_5. The cycle type distributions for these groups are as follows:

cycle type :	1	2	(2, 2)	3	(2, 3)	4	5
Z_5	1						4
D_{10}	1		5				4
F_{20}	1		5			10	4
A_5	1		15	20			24
S_5	1	10	15	20	20	30	24.

Given this information, the irreducibility of $x^5 - x - 1$ (giving the transitivity on the 5 roots) and the cycle type (2,3) immediately shows that the Galois group of $x^5 - x - 1$ is S_5.

Consider now the polynomial $x^5 + 15x + 12$. The discriminant is $2^{10}3^4 5^5$ so the Galois group is not contained in A_5. There are two possibilities: S_5 or F_{20}. One can easily determine which is more likely by factoring the polynomial modulo a number of small primes and comparing the distribution of cycle types with those in the table above. This does not *prove* the probable Galois group is actually correct. To decide which of S_5 and F_{20} is correct one can compute the resolvent polynomial $R(x)$ of degree 15 whose roots are the distinct permutations under S_5 of $(\alpha_1 + \alpha_2 - \alpha_3 - \alpha_4)^2$ for 4 of the roots $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ of $f(x)$. By definition, S_5 is transitive on the roots of $R(x)$ and it is not difficult to check using the explicit generators for F_{20} given above that F_{20} is not transitive on these 15 values. It follows that $R(x)$ will be a reducible polynomial over $\mathbb{Q}$ if and only if the Galois group of the quintic is F_{20}. One finds that for $x^5 + 15x + 12$ the resolvent polynomial $R(x)$ factors into a polynomial of degree 5 and a polynomial of degree 10, hence the Galois group for this quintic is F_{20}. One can also use Exercise 21 of the previous section (cf. Exercise 6), which is also based on the computation of a related resolvent polynomial.

(2) Consider the polynomial $x^7 - 14x^5 + 56x^3 - 56x + 22$. The discriminant is computed to be $2^6 7^{10}$ so the Galois group is contained in A_7.

Factoring the polynomial for the 42 primes not equal to 7 between 3 and 193 gives a cycle type distribution of 1 1-cycle (2.38 %), 30 (3,3)-cycles (71.43 %), 11 7-cycles (26.19 %). There are 7 isomorphism classes of transitive subgroups of S_7, 4 of them contained in A_7. Of these, one contains no (3,3)-cycles, which leaves the three possibilities A_7, $GL_3(\mathbb{F}_2)$, or F_{21}, the Frobenius group of order 21 (which has generators (1 2 3 4 5 6 7) and (2 3 5)(4 7 6) in S_7). The cycle type distributions for these three are as follows:

cycle type:	1	2	(2, 2)	3	(2, 2, 3)	(3, 3)	(2, 4)	5	7
F_{21}	1					14			6
$GL_3(\mathbb{F}_2)$	1		21			56	42		48
A_7	1	21	105	70	210	280	630	504	720

It follows that there is a strong probability that the Galois group of this polynomial is the Frobenius group of order 21. This is actually the case (the verification requires computation of a resolvent of degree 35 and factoring it over $\mathbb{Z}$ — there are three factors, of degrees 7,7, and 21).

EXERCISES

1. Let p be a prime. Prove that the polynomial $x^4 + 1$ splits mod p either into two irreducible quadratics or into 4 linear factors using Corollary 41 together with the knowledge that the Galois group of this polynomial is the Klein 4-group.
2. (Cf. Exercise 48 of Section 6).
   (a) Let K be the splitting field of $x^6 - 2x^3 - 2$. Prove that if $[K : \mathbb{Q}] = 12$ then $K = \mathbb{Q}(\sqrt[3]{2}, i, \sqrt{3})$ and K is generated over the biquadratic field $F = \mathbb{Q}(i, \sqrt{3})$ by $\alpha = \sqrt[3]{1 + \sqrt{3}}$ and by $\beta = \sqrt[3]{1 - \sqrt{3}}$. Show that if this is the case then the elements of order 3 in $\mathrm{Gal}(K/\mathbb{Q})$ lie in $\mathrm{Gal}(K/F)$. Conclude that any element of $\mathrm{Gal}(K/\mathbb{Q})$ of order 3 maps α to another cube root of $1 + \sqrt{3}$ and maps β to another cube root of $1 - \sqrt{3}$ and if it is the identity on α or β then it is the identity on all of K.
   (b) Show that the factorization of $f(x)$ into irreducibles over $\mathbb{F}_{13}$ is the polynomial $(x - 7)(x - 8)(x - 11)(x^3 + 3)$ and use Corollary 41 to show that $[K : \mathbb{Q}] = 36$.
   (c) Knowing that $G = \mathrm{Gal}(K/\mathbb{Q})$ is of order 36 determine all the elements of G explicitly and in particular show that G is isomorphic to $S_3 \times S_3$.
3. Prove that the Galois group of $x^5 + 20x + 16$ is A_5.
4. Prove that the Galois group of $x^5 + x^4 - 4x^3 - 3x^2 + 3x + 1$ is cyclic of order 5. [Show this is the minimal polynomial of $\zeta_{11} + \zeta_{11}^{-1}$.]
5. Prove that the Galois group of $x^5 + 11x + 44$ is the dihedral group D_{10} (cf. Exercise 21 of Section 7).
6. Prove that the Galois group of $x^5 + 15x + 12$ is F_{20}, the Frobenius group of order 20 (cf. Exercise 21 of Section 7).
7. Prove that the Galois group of $x^6 + 24x - 20$ is A_6.
8. Prove that the Galois group of $x^7 + 7x^4 + 14x + 3$ is A_7.

9. Determine a polynomial of degree 7 whose Galois group is cyclic of order 7.

10. Determine the probable Galois group of $x^7 - 7x + 3$.

14.9 TRANSCENDENTAL EXTENSIONS, INSEPARABLE EXTENSIONS, INFINITE GALOIS GROUPS

This section collects some results on arbitrary extensions E/F. These results supplement those of the preceding sections and complete the basic picture of how an arbitrary (possibly infinite) extension decomposes. Since this section is primarily intended as a survey, none of the proofs are included; whenever these proofs can be easily supplied by the reader we indicate this either in the text or (with hints) in the exercises.

Throughout this section E/F is an extension of fields. Recall that an element of E which is not algebraic over F is called transcendental over F. Keep in mind that extensions involving transcendentals are always of infinite degree. We generally reserve the expression "t is an 'indeterminate' over F", when we are thinking of evaluating t. Field theoretically, however, the terms transcendental and indeterminate are synonymous (so that the subfield $\mathbb{Q}(\pi)$ of $\mathbb{R}$ and the field $\mathbb{Q}(t)$ are isomorphic).

Definition.

(1) A subset $\{a_1, a_2, \dots, a_n\}$ of E is called *algebraically independent* over F if there is no nonzero polynomial $f(x_1, x_2, \dots, x_n) \in F[x_1, x_2, \dots, x_n]$ such that $f(a_1, a_2, \dots, a_n) = 0$. An arbitrary subset S of E is called *algebraically independent* over F if every finite subset of S is algebraically independent. The elements of S are called *independent transcendentals* over F.

(2) A *transcendence base* for E/F is a maximal subset (with respect to inclusion) of E which is algebraically independent over F.

Note that if E/F is algebraic, the empty set is the only algebraically independent subset of E. In particular, elements of an algebraically independent set are necessarily transcendental. Moreover, one easily checks that $S \subseteq E$ is an algebraically independent set over F if and only if each $s \in S$ is transcendental over $F(S - \{s\})$. It is also an easy exercise to see that S is a transcendence base for E/F if and only if S is a set of algebraically independent transcendentals over F and E is algebraic over $F(S)$.

Theorem. The extension E/F has a transcendence base and any two transcendence bases of E/F have the same cardinality.

Proof: The first statement is a standard Zorn's Lemma argument. The proof of the second uses the same "Replacement Lemma" idea as was used to prove that any two bases of a vector space have the same cardinality.

Definition. The cardinality of a transcendence base for E/F is called the *transcendence degree* of E/F.

Algebraic extensions are precisely the extensions of transcendence degree 0.

One special case of this theorem is when E is *finitely generated* over F, that is, $E = F(\alpha_1, \alpha_2, \dots, \alpha_n)$, for some (not necessarily algebraically independent) elements $\alpha_1, \dots, \alpha_n$ of E. It is clear that we may renumber $\alpha_1, \dots, \alpha_n$ so that $\alpha_1, \dots, \alpha_m$ are independent transcendentals and $\alpha_{m+1}, \dots, \alpha_n$ are algebraic over $F(\alpha_1, \dots, \alpha_m)$ (so E is a finite extension of the latter field). In this case E is called a *function field in m variables* over F. Such fields play a fundamental role in algebraic geometry as fields of functions on m-dimensional surfaces. For instance, when $F = \mathbb{C}$ and $m = 1$, these fields arise in analysis as fields of meromorphic functions on compact Riemann surfaces.

Note that if S_1 and S_2 are transcendence bases for E/F it is not necessarily the case that $F(S_1) = F(S_2)$. For example, if t is transcendental over $\mathbb{Q}$, $\{t\}$ and $\{t^2\}$ are both transcendence bases for $\mathbb{Q}(t)/\mathbb{Q}$ but (as we shall see shortly) $\mathbb{Q}(t^2)$ is a proper subfield of $\mathbb{Q}(t)$.

We now see that if $x_1, x_2, \dots, x_n$ are indeterminates over F and

$$f(x) = (x - x_1)(x - x_2)\cdots(x - x_n) \tag{14.28}$$

is the general polynomial of degree n, then the set of n elementary symmetric functions $s_1, s_2, \dots, s_n$ in the x_i's are also independent transcendentals over F. This is because $x_1, \dots, x_n$ is a transcendence base for $E = F(x_1, \dots, x_n)$ over F (so the transcendence degree is n) and E is algebraic over $F(s_1, \dots, s_n)$ (of degree $n!$). The theorem forces $s_1, \dots, s_n$ to be a transcendence base for this extension as well (in particular, they are independent transcendentals). The general polynomial of degree n over F may therefore equivalently be defined by taking $a_1, \dots, a_n$ to be any independent transcendentals (or indeterminates) and letting

$$f(x) = x^n + a_1 x^{n-1} + \cdots + a_n \tag{14.29}$$

where the roots of f are denoted by $x_1, \dots, x_n$ (and $s_i = (-1)^i a_i$).

Definition. An extension E/F is called *purely transcendental* if it has a transcendence base S such that $E = F(S)$.

In the preceding discussion, both $F(x_1, \dots, x_n)$ and $F(s_1, \dots, s_n)$ are purely transcendental over F. As an exercise (following) one can show that $\mathbb{Q}(t, \sqrt{t^3 - t}\,)$ is not a purely transcendental extension of $\mathbb{Q}$ even though it contains no elements that are algebraic over $\mathbb{Q}$ other than those in $\mathbb{Q}$ itself (i.e., the process of decomposing a general extension into a purely transcendental extension followed by an algebraic extension cannot generally be reversed so that the algebraic piece occurs first).

If E is a purely transcendental extension of F of transcendence degree $n = 1$ or 2 and L is an intermediate field, $F \subseteq L \subseteq E$ with the same transcendence degree, then L is again a purely transcendental extension of F (Lüroth ($n = 1$), Castelnuovo ($n = 2$)). This result is not true if the transcendence degree is ≥ 3, however, although examples where L fails to be purely transcendental are difficult to construct. For extensions of transcendence degree 1 the intermediate fields are described by the following theorem.

Theorem. Let t be transcendental over F.

(1) (Lüroth) If $F \subseteq K \subseteq F(t)$, then $K = F(r)$, for some $r \in F(t)$. In particular, every nontrivial extension of F contained in $F(t)$ is purely transcendental over F.

(2) If $P = P(t)$, $Q = Q(t)$ are nonzero relatively prime polynomials in $F[t]$ which are not both constant,

$$[F(t) : F(P/Q)] = \max(\deg P, \deg Q).$$

Proof: The proof of (2) is outlined in Exercise 18 of Section 13.2.

By part (2) of this theorem we see that $F(P/Q) = F(t)$ if and only if P, Q are nonzero relatively prime polynomials of degree ≤ 1 (not both constant). Thus $F(r) = F(t)$ if and only if $r = \dfrac{at+b}{ct+d}$, where $a, b, c, d \in F$ and $ad - bc \neq 0$ (called a *fractional linear transformation of* t). For *any* $r \in F(t) - F$ the map $t \mapsto r$ extends to an embedding of $F(t)$ into itself which is the identity on F. This embedding is surjective (i.e., is an automorphism of $F(t)$) precisely for the fractional linear transformations. Furthermore, the map

$$GL_2(F) \to \mathrm{Aut}(F(t)/F) \quad \text{defined by} \quad A = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \mapsto \sigma_A,$$

where σ_A denotes the automorphism of $F(t)$ defined by mapping t to $(at+b)\big/(ct+d)$, is a surjective homomorphism with kernel consisting of the scalar matrices. Thus

$$\mathrm{Aut}(F(t)/F) \cong PGL_2(F)$$

where $PGL_2(F) = GL_2(F)/\{\lambda I \mid \lambda \in F^\times\}$ gives the group of automorphisms of this transcendental extension (cf. Exercise 8 of Section 1).

When $\mathbb{F}$ is a finite field of order q, $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F}) \cong PGL_2(\mathbb{F})$ is a finite group of order $q(q-1)(q+1)$. By Corollary 11 if K is the fixed field of $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F})$, then $\mathbb{F}(t)$ is Galois over K with Galois group equal to $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F})$. In particular, the fixed field of $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F})$ is not $\mathbb{F}$ in this case.

This also provides further examples of the Galois correspondence which can be written out completely for small values of q. For instance, if $q = |\mathbb{F}| = 2$, $PGL_2(\mathbb{F})$ is nonabelian of order 6, hence is isomorphic to S_3, and has the following lattice of subgroups:

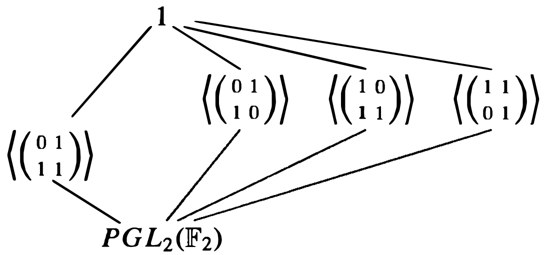

1
$\langle\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\rangle$
$\langle\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\rangle$
$\langle\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\rangle$
$\langle\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}\rangle$
$PGL_2(\mathbb{F}_2)$

Fig. 5

The field $\mathbb{F}(t)$ is of degree 6 over the fixed field K of $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F})$ and the lattice of subfields $K \subseteq L \subseteq \mathbb{F}(t)$ is dual to the lattice of subgroups of S_3. The fixed field of a cyclic subgroup $\langle \sigma \rangle$ is easily found (via the preceding theorem) by finding a rational function r in t which is fixed by σ such that $[\mathbb{F}(t) : \mathbb{F}(r)] = |\sigma|$. For example, if $\sigma : t \mapsto 1/(1+t)$, then σ has order 3. The rational function

$$r = t + \sigma(t) + \sigma^2(t) = \frac{t^3+t+1}{t(t+1)}$$

is fixed by σ and $[\mathbb{F}(t) : \mathbb{F}(r)] = 3$ (by part (2) of the theorem). Since $\mathbb{F}(r)$ is contained in the fixed field of $\langle \sigma \rangle$ and the degree of $\mathbb{F}(t)$ over the fixed field is 3, $\mathbb{F}(r)$ is the fixed field of $\langle \sigma \rangle$. In this way one can explicitly describe the lattice of all subfields of $\mathbb{F}(t)$ containing K shown in Figure 6.

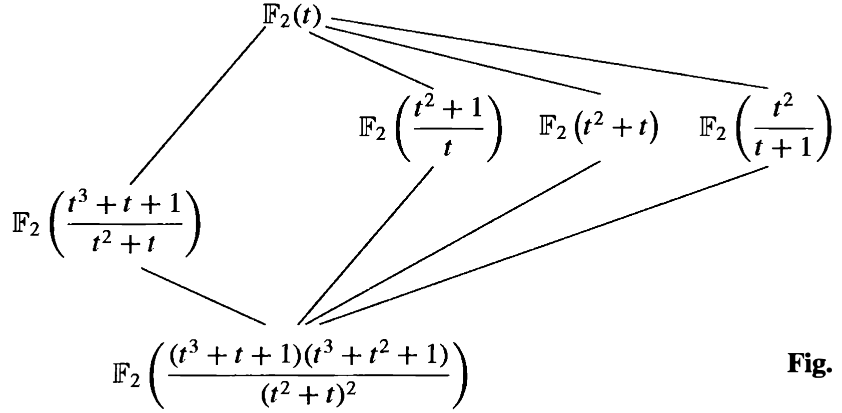

$\mathbb{F}_2(t)$
$\mathbb{F}_2\left(\frac{t^2+1}{t}\right)$
$\mathbb{F}_2\left(t^2+t\right)$
$\mathbb{F}_2\left(\frac{t^2}{t+1}\right)$
$\mathbb{F}_2\left(\frac{t^3+t+1}{t^2+t}\right)$
$\mathbb{F}_2\left(\frac{(t^3+t+1)(t^3+t^2+1)}{(t^2+t)^2}\right)$

Fig. 6

Purely transcendental extensions of $\mathbb{Q}$ play an important role in the problem of realizing finite groups as Galois groups over $\mathbb{Q}$. We describe a deep result of Hilbert which is fundamental to this area of research. If $a_1, a_2, \ldots, a_n$ are independent indeterminates over a field F, we may evaluate (or *specialize*) $a_1, \ldots, a_n$ at any elements of F, i.e., substitute values in F for the "variables" $a_1, a_2, \ldots, a_n$. If E is a Galois extension of $F(a_1, \ldots, a_n)$, then E is obtained as a splitting field of a polynomial whose coefficients lie in $F[a_1, \ldots, a_n]$. Any specialization of $a_1, \ldots, a_n$ into F maps this polynomial into one whose coefficients lie in F. The specialization of E is the splitting field of the resulting specialized polynomial.

Theorem. (Hilbert) Let $x_1, x_2, \ldots, x_n$ be independent transcendentals over $\mathbb{Q}$, let $E = \mathbb{Q}(x_1, \ldots, x_n)$ and let G be a finite group of automorphisms of E with fixed field K. If K is a purely transcendental extension of $\mathbb{Q}$ with transcendence basis $a_1, a_2, \ldots, a_n$, then there are infinitely many specializations of $a_1, \ldots, a_n$ in $\mathbb{Q}$ such that E specializes to a Galois extension of $\mathbb{Q}$ with Galois group isomorphic to G.

Hilbert's Theorem gives a sufficient condition for the specialized extension not to collapse. In general, the Galois group of the specialized extension is a subgroup of G (cf. Proposition 19) and may be a proper subgroup of G. It is also known that the fixed field K need not always be a purely transcendental extension of $\mathbb{Q}$. An example of this occurs when G is the cyclic group of order 47.

This theorem can be used to give another proof of Proposition 42:

Corollary. S_n is a Galois group over $\mathbb{Q}$, for all n.

Proof of the Corollary: We have already proved that the fixed field of S_n acting in the obvious fashion on $\mathbb{Q}(x_1, \dots, x_n)$ is purely transcendental over $\mathbb{Q}$ (with the elementary symmetric functions as a transcendence base), so Hilbert's Theorem immediately implies the corollary.

The hypothesis that K be purely transcendental over $\mathbb{Q}$ is crucial to the proof of Hilbert's Theorem. Every finite group is isomorphic to a subgroup of S_n and so acts on $\mathbb{Q}(x_1, \dots, x_n)$ for some n. It is not known, however, even for the subgroup A_n of S_n whether its fixed field under the obvious action is a purely transcendental extension of $\mathbb{Q}$ (although it is known by other means that A_n is a Galois group over $\mathbb{Q}$ for all n). Thus there are a number of important open problems in this area of research.

One should also notice that Hilbert's Theorem does not work when the base field $\mathbb{Q}$ is replaced by an arbitrary field F (suppose F were algebraically closed, for instance). In particular, as noted earlier, the general polynomial $f(x)$ in Section 6 has Galois group S_n over $F(a_1, \dots, a_n)$ for any F, but when F is a finite field, the specialized extension obtained from its splitting field is always cyclic.

We next expand on the theory of inseparable extensions described in Section 13.5. Let p be a prime and let F be a field of characteristic p.

Definition. An algebraic extension E/F is called *purely inseparable* if for each $\alpha \in E$ the minimal polynomial of α over F has only one distinct root.

It is easy to see that the following are equivalent:

(1) E/F is purely inseparable
(2) if $\alpha \in E$ is separable over F, then $\alpha \in F$
(3) if $\alpha \in E$, then $\alpha^{p^n} \in F$ for some n (depending on α), and $m_{\alpha,F}(x) = x^{p^n} - \alpha^{p^n}$.

The following easy proposition describes composites of separable and purely inseparable extensions.

Proposition. If E_1 and E_2 are subfields of E which are both separable (or both purely inseparable) extensions of F, then their composite E_1E_2 is separable (purely inseparable, respectively) over F.

Proof: Exercise.

One immediate consequence of this is the following result.

Proposition. Let E/F be an algebraic extension. Then there is a unique field E_{sep} with $F \subseteq E_{sep} \subseteq E$ such that E_{sep} is separable over F and E is purely inseparable over E_{sep}. The field E_{sep} is the set of elements of E which are separable over F.

The degree of E_{sep}/F is called the *separable degree* of E/F and the degree of E/E_{sep} is called the *inseparable degree* of E/F (often denoted as $[E : F]_s$ and $[E : F]_i$ respectively). The product of these two degrees is the (ordinary) degree. The propositions immediately give the following corollary.

Corollary. Separable degrees (respectively inseparable degrees) are multiplicative.

When E is generated over F by the root of an irreducible polynomial $p(x) \in F[x]$ the separable and inseparable degrees of the extension E/F are the same as the separable and inseparable degrees of the polynomial $p(x)$ defined in Section 13.5.

The proposition asserts that any algebraic extension may be decomposed into a separable extension followed by a purely inseparable one. Exercise 3 at the end of this section outlines an example illustrating that this decomposition cannot generally be reversed, namely an extension which is not a separable extension of a purely inseparable extension. We shall shortly state conditions on an extension under which the decomposition into separable and purely inseparable subextensions may be reversed.

We now know that an arbitrary extension E/F can be decomposed into a purely transcendental extension $F(S)$ of F followed by a separable extension E_1 of $F(S)$ followed by a purely inseparable extension E/E_1. In certain instances the inseparability in the algebraic extension at the "top" may be removed by a judicious choice of transcendence base:

Proposition. If E is a finitely generated extension of a perfect field F, then there is a transcendence base T of E/F such that E is a separable (algebraic) extension of $F(T)$.

A transcendence base T as described in the proposition is called a *separating transcendence base.* Exercise 4 at the end of this section illustrates this with a nontrivial example.

Recall that an extension E/F is *normal* if it is the splitting field of some (possibly infinite) set of polynomials in $F[x]$ (in particular, normal extensions are algebraic but not necessarily finite or separable). We previously used the synonymous term splitting field and the term normal is reintroduced here in the context of arbitrary algebraic extensions since it is used frequently in the literature, often in the context of embeddings of a field into an algebraic closure. Although the following set of equivalences can be gleaned from the preceding sections, the reader should write out a complete proof, checking that the arguments work for both infinite and inseparable extensions:

Proposition. Let E/F be an arbitrary algebraic extension and let Ω be an algebraic closure of E. The following are equivalent:

(1) E/F is a normal extension (i.e., is the splitting field over F of some set of polynomials in $F[x]$)

(2) whenever $\sigma : E \to \Omega$ is an embedding such that $\sigma|_F$ is the identity, $\sigma(E) = E$

(3) whenever an irreducible polynomial $f(x) \in F[x]$ has one root in E, it has all its roots in E.

In general, any embedding of a normal extension E/F into an algebraic closure of E which extends the identity embedding of F is an automorphism of E, i.e., is an element of $\text{Aut}(E/F)$. Moreover, the number of such automorphisms equals the separable degree of E/F, provided the latter is finite:

$$\text{if } E/F \text{ is a normal extension and } [E : F]_s \text{ is finite,} \quad |\text{Aut}(E/F)| = [E : F]_s.$$

If $[E : F]_s$ is infinite we shall see shortly that $|\text{Aut}(E/F)|$ is also infinite but need not be of the same cardinality.

If E/F is a normal extension whose separable degree is finite, let E_0 be the fixed field of $\text{Aut}(E/F)$. By Corollary 11, E/E_0 is a (separable) Galois extension whose degree equals $|\text{Aut}(E/F)|$. It follows that E_0/F must be purely inseparable (of degree equal to $[E : F]_i$), i.e., the separable and purely inseparable pieces of the extension may be reversed for normal extensions. More precisely, we easily obtain the following proposition.

Proposition. If E/F is normal with $[E : F]_s < \infty$, then $E = E_{sep}E_{pi}$, where E_{pi} is a purely inseparable extension of F (E_{pi} consists of all purely inseparable elements of E over F) and $E_{sep} \cap E_{pi} = F$.

Finally, we mention how Galois Theory generalizes to infinite extensions.

Definition. An extension E/F is called *Galois* if it is algebraic, normal and separable. In this case $\text{Aut}(E/F)$ is called the *Galois group* of the extension and is denoted by $\text{Gal}(E/F)$.

For infinite extensions there need not be a bijection between the set of all subgroups of the Galois group and the set of all subfields of E containing F, as the following example illustrates.

Let E be the subfield of $\mathbb{R}$ obtained by adjoining to $\mathbb{Q}$ all square roots of positive rational numbers. One easily sees that E may also be described as the splitting field of the set of polynomials $x^2 - p$, where p runs over all primes in $\mathbb{Z}^+$. Note that E is a (countably) infinite Galois extension of $\mathbb{Q}$. Since every automorphism σ of E is determined by its action on the square roots of the primes and σ either fixes or negates each of these, σ^2 is the identity automorphism. It follows that $\text{Aut}(E)$ is an infinite elementary abelian 2-group. Thus $\text{Aut}(E)$ is an infinite dimensional vector space over $\mathbb{F}_2$. By an exercise in the section on dual spaces (Section 11.3) the number of nonzero homomorphisms of $\text{Aut}(E)$ into $\mathbb{F}_2$ is uncountable, whence their kernels (which are subspaces of co-dimension 1) are uncountable in number (and distinct). Thus $\text{Aut}(E)$ has *uncountably* many subgroups of index 2, whereas $\mathbb{Q}$ has only a *countable* number of quadratic extensions.

The basic problem is that many (most) subgroups of $\text{Gal}(E/F)$ do not correspond (in a bijective fashion) to subfields of E containing F. In order to pick out the "right" set of subgroups of $\mathrm{Gal}(E/F)$ we must introduce a topology on this group (called the Krull topology). The axioms for the collection of (topologically) closed subsets of a topological space are precisely the bookkeeping devices which single out the relevant subgroups (these are listed in Section 15.2). Galois theory for finite extensions force certain subgroups of finite index to be closed sets and these in turn determine the topology on the entire group (as we might expect since every extension of F inside E is a composite of finite extensions). Moreover, the Galois group of E/F is the inverse limit of the collection of finite groups $\mathrm{Gal}(K/F)$, where K runs over all finite Galois extensions of F contained in E (cf. Exercise 10, Section 7.6).

Theorem. (Krull) Let E/F be a Galois extension with Galois group G. Topologize G by taking as a base for the closed sets the subgroups of G which are the fixing subgroups of the finite extensions of F in E, together with all left and right cosets of these subgroups. Then with this ("Krull") topology the closed subgroups of G correspond bijectively with the subfields of E containing F and the corresponding lattices are dual. Closed normal subgroups of G correspond to normal extensions of F in E.

One important area of current research is to describe (as a topological group) the Galois group of certain field extensions such as $\overline{F}/F$, where $\overline{F}$ is the algebraic closure of F. Little is known about the latter group when $F = \mathbb{Q}$ (in particular, its normal subgroups of finite index, i.e., which finite groups occur as Galois groups over $\mathbb{Q}$, are not known). If E is the algebraic closure of the finite field $\mathbb{F}_p$, the Galois group of this extension is the topologically cyclic group $\widehat{\mathbb{Z}}$ with the Frobenius automorphism as a topological generator. The group $\widehat{\mathbb{Z}}$ is an uncountable group (in particular, is not isomorphic to $\mathbb{Z}$) with the property that every closed subgroup of finite index is normal with cyclic quotient. Note that $\widehat{\mathbb{Z}}$ must also have nontrivial infinite closed subgroups (unlike $\mathbb{Z}$) since E contains proper subfields which are infinite over $\mathbb{F}_p$ (such as the composite of all extensions of $\mathbb{F}_p$ of q-power degree, for any prime q — this Galois extension of $\mathbb{F}_p$ has Galois group $\mathbb{Z}_q$, the q-adic integers, as described in Exercise 11 of Section 7.6).

EXERCISES

1. Prove that every purely inseparable extension is normal.
2. Let p be a prime and let $K = \mathbb{F}_p(x, y)$ with x and y independent transcendentals over $\mathbb{F}_p$. Let $F = \mathbb{F}_p(x^p - x, y^p - x)$.
   (a) Prove that $[K : F] = p^2$ and the separable degree and inseparable degree of K/F are both equal to p.
   (b) Prove that there is a subfield E of K containing F which is purely inseparable over F of degree p (so then K is a separable extension of E of degree p). [Let $s = x^p - x \in F$ and $t = y^p - x \in F$ and consider $s - t$.]
3. Let p be an odd prime, let s and t be independent transcendentals over $\mathbb{F}_p$, and let F be the field $\mathbb{F}_p(s, t)$. Let β be a root of $x^2 - sx + t = 0$ and let α be a root of $x^p - \beta = 0$ (in some algebraic closure of F). Set $E = F(\beta)$ and $K = F(\alpha)$.
   (a) Prove that E is a Galois extension of F of degree 2 and that K is a purely inseparable extension of E of degree p.

(b) Prove that K is not a normal extension of F. [If it were, conjugate β over F to show that K would contain a p^{th} root of s and then also a p^{th} root of t, so $[K : F] \geq p^2$, a contradiction.]

(c) Prove that there is no field K_0 such that $F \subseteq K_0 \subseteq K$ with K_0/F purely inseparable and K/K_0 separable. [If there were such a field, use Exercise 1 and the fact that the composite of two normal extensions is again normal to show that K would be a normal extension of F.]

4. Under the notation of the previous exercise prove that α, s is a separating transcendence base for K over $\mathbb{F}_p$.

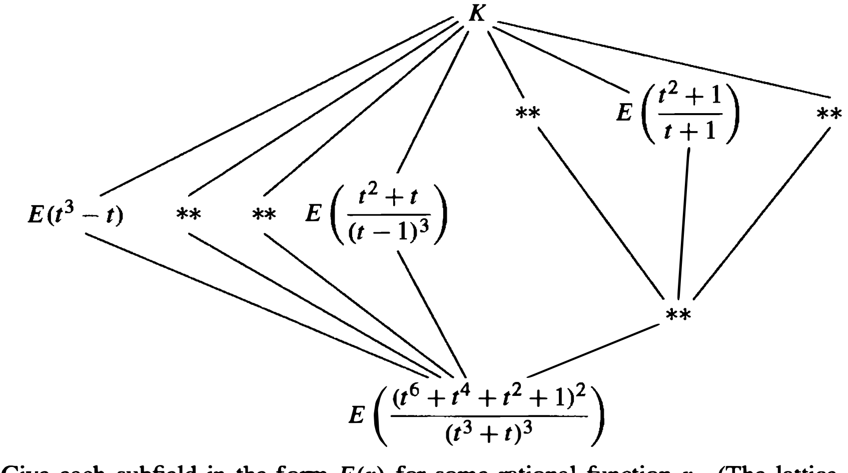
5. Let p be a prime, let t be transcendental over $\mathbb{F}_p$ and let K be obtained by adjoining to $\mathbb{F}_p(t)$ all p-power roots of t. Prove that K has transcendence degree 1 over $\mathbb{F}_p$ and has no separating transcendence base.

6. Show that if t is transcendental over $\mathbb{Q}$ then $\mathbb{Q}(t, \sqrt{t^3 - t}\,)$ is not a purely transcendental extension of $\mathbb{Q}$. (This is an example of what is called an *elliptic* function field.)

7. Let k be the field with 4 elements, t a transcendental over k, $F = k(t^4 + t)$ and $K = k(t)$.
   (a) Show that $[K : F] = 4$.
   (b) Show that K is separable over F.
   (c) Show that K is Galois over F.
   (d) Describe the lattice of subgroups of the Galois group and the corresponding lattice of subfields of K, giving each subfield in the form $k(r)$, for some rational function r.

8. Let p be an odd prime, k an algebraically closed field of characteristic p and let t be transcendental over k. Suppose F is a degree 2 field extension of $k(t)$. Show that F can be written in the form $k(t, y)$, for some $y \in F$ with $y^2 \in k(t)$ and y transcendental over k. If $y^2 = 4t^3 - t - 1$, find $[F : k(y)]$ and describe $k(t) \cap k(y)$ as $k(r)$, for some $r \in k(t)$.

9. Let t be transcendental over $\mathbb{F}_3$, let $K = \mathbb{F}_3(t)$, let $G = \text{Aut}(K/\mathbb{F}_3)$ and let F be the fixed field of G.
   (a) Prove $G \cong S_4$ and deduce that there is a unique field E with $F \subseteq E \subseteq K$ and $[E : F] = 2$. [Recall that $G \cong PGL_2(\mathbb{F}_3)$; show that $GL_2(\mathbb{F}_3)$ permutes the 4 lines in a 2-dimensional vector space over $\mathbb{F}_3$ and the kernel of this permutation representation is the scalar matrices.]
   (b) Complete the description of the lattice of subfields of K containing E:



K
$**$
$E\left(\frac{t^2+1}{t+1}\right)$
$**$
$E(t^3 - t)$
$**$
$**$
$E\left(\frac{t^2+t}{(t-1)^3}\right)$
$**$
$E\left(\frac{(t^6+t^4+t^2+1)^2}{(t^3+t)^3}\right)$

Give each subfield in the form $E(r)$ for some rational function r. (The lattice of subgroups of A_4 appears in Section 3.5).

10. Prove that a purely transcendental proper extension of a field is never algebraically closed.

11. Let S be a set of independent transcendentals over a field F and let Ω be an algebraic closure of $F(S)$. Prove that any permutation on S extends to an element of $\mathrm{Aut}(F(S)/F)$. Prove that any such automorphism of $F(S)$ extends to an automorphism of Ω. Deduce that $\mathbb{C}$ has infinitely many automorphisms.

12. Let K be a subfield of $\mathbb{C}$ maximal with respect to the property "$\sqrt{2} \notin K$."

(a) Show such a field K exists.

(b) Show that $\mathbb{C}$ is algebraic over K.

(c) Prove that every finite extension of K in $\mathbb{C}$ is Galois with Galois group a cyclic 2-group.

(d) Deduce that $[\mathbb{C} : K]$ is countable (and not finite).

13. Let K be the fixed field in $\mathbb{C}$ of an automorphism of $\mathbb{C}$. Prove that every finite extension of K in $\mathbb{C}$ is cyclic.

14. Let K_n be the splitting field of $(x^2 - p_1)(x^2 - p_2)\cdots(x^2 - p_n)$ over $\mathbb{Q}$, where $p_1, \ldots, p_n$ are the first n primes. Prove that the Galois group of $K_n/\mathbb{Q}$ is an elementary abelian 2-group of order 2^n.

15. Let $K_0 = \mathbb{Q}$ and for $n \geq 0$ define the field K_{n+1} as the extension of K_n obtained by adjoining to K_n all roots of all cubic polynomials over K_n. Let K be the union of the subfields K_n, $n \geq 0$. Prove that K is a Galois extension of $\mathbb{Q}$. Prove that every cubic polynomial over K splits completely over K. Prove that there are nontrivial algebraic extensions of K.

16. Let F be the composite of all the splitting fields of irreducible cubics over $\mathbb{Q}$. Prove that F does not contain all quadratic extensions of $\mathbb{Q}$.

17. Let $K_0 = \mathbb{Q}$ and for $n \geq 0$ define the field K_{n+1} as the extension of K_n obtained by adjoining to K_n all radicals of elements in K_n. Let K be the union of the subfields K_n, $n \geq 0$. Prove that K is a Galois extension of $\mathbb{Q}$. Prove that there are no nontrivial solvable Galois extensions of K. Prove that there are nontrivial Galois extensions of K.

18. Let $F_0 = \mathbb{Q}$ and for $n \geq 0$ define the field F_{n+1} as the extension of F_n obtained by adjoining to F_n all real radicals of elements in F_n. Let F be the union of the subfields F_n, $n \geq 0$. Let K^+ be the fixed field of complex conjugation restricted to the field K in the previous exercise (the maximal real subfield of K). Prove that $F \neq K^+$.

19. This exercise proves that if K/F is a Galois extension of fields, then $\mathrm{Gal}(K/F)$ is isomorphic to $\varprojlim \mathrm{Gal}(L/F)$, where the inverse limit is taken over all the finite Galois extensions L of F contained in K.

(a) Show that K is the union of the fields L.

(b) Prove that the map $\varphi : \mathrm{Gal}(K/F) \to \varprojlim \mathrm{Gal}(L/F)$ defined by mapping σ in $\mathrm{Gal}(K/F)$ to $(\ldots, \sigma|_L, \ldots)$, where $\sigma|_L$ is the restriction of σ to L, is a homomorphism.

(c) Show that φ is injective.

(d) If $(\ldots, \sigma_L, \ldots) \in \varprojlim \mathrm{Gal}(L/F)$, define $\sigma \in \mathrm{Gal}(K/F)$ by $\sigma(\alpha) = \sigma_L(\alpha)$ if $\alpha \in L$. Prove that σ is a well defined automorphism and deduce that φ is surjective.

Part V

INTRODUCTION TO COMMUTATIVE RINGS, ALGEBRAIC GEOMETRY, AND HOMOLOGICAL ALGEBRA

In this part of the book we continue the study of rings and modules, concentrating first on commutative rings. The topic of commutative algebra, which is of interest in its own right, is also a basic foundation for other areas of algebra. To indicate some of the importance of the algebraic topics introduced, we parallel the development of the ring theory in Chapter 15 with an introduction to affine algebraic geometry. Each section first presents the basic algebraic theory and then follows with an application of those ideas to geometry together with an indication of computational methods using the theory of Gröbner bases from Chapter 9. The purpose here is twofold: the first is to present an application of algebraic techniques in the important branch of mathematics called Algebraic Geometry, and the second is to indicate some of the motivations for the algebraic concepts introduced from their origins in geometric questions.

This connection of geometry and algebra shows a rich interplay between these two areas of mathematics and demonstrates again how results and structures in one circle of mathematical ideas provide insights into another.

In Chapter 16 we continue with some of the fundamental structures involving commutative rings, culminating with Dedekind Domains and a structure theorem for modules over such rings which is a generalization of the structure theorem for modules over P.I.D.s in Chapter 12.

In Chapter 17 we describe some of the basic techniques of "homological algebra," which continues with some of the questions raised by the failure of exactness of some of the sequences considered in Chapter 10. The cohomology of groups in this chapter is intended to serve both as a more in-depth application of homological algebra to see its uses in practice, and as a relatively self contained exposition of this important topic.

CHAPTER 15

Commutative Rings and Algebraic Geometry

Throughout this chapter R will denote a commutative ring with $1 \neq 0$.

15.1 NOETHERIAN RINGS AND AFFINE ALGEBRAIC SETS

In this section we study Noetherian rings in greater detail. These are a natural generalization of Principal Ideal Domains and were introduced briefly in Chapter 12. Note that when R is considered as a left module over itself, its R-submodules are precisely its ideals, so the definition in Section 1 of Chapter 12 may be phrased in the following form:

Definition. A commutative ring R is said to be *Noetherian* or to satisfy the *ascending chain condition on ideals* (or *A.C.C. on ideals*) if there is no infinite increasing chain of ideals in R, i.e., whenever $I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots$ is an increasing chain of ideals of R, then there is a positive integer m such that $I_k = I_m$ for all $k \geq m$.

Proposition 1. If I is an ideal of the Noetherian ring R, then the quotient R/I is a Noetherian ring. Any homomorphic image of a Noetherian ring is Noetherian.

Proof: If R is a ring and I is an ideal in R, then any infinite ascending chain of ideals in the quotient R/I would correspond by the Lattice Isomorphism Theorem to an infinite ascending chain of ideals in R. This gives the first statement, and the second follows by the first Isomorphism Theorem.

Theorem 2. The following are equivalent:

(1) R is a Noetherian ring.
(2) Every nonempty set of ideals of R contains a maximal element under inclusion.
(3) Every ideal of R is finitely generated.

Proof: The proof is identical to that of Theorem 1 in Section 12.1 in the special case where the R-module M is R itself (and submodules are ideals).

Examples

Every Principal Ideal Domain is Noetherian since it satisfies condition (3) of Theorem 2. In particular, $\mathbb{Z}$, the polynomial ring $k[x]$ where k is a field, and the Gaussian integers $\mathbb{Z}[i]$, are Noetherian rings. The ring $\mathbb{Z}[x_1, x_2, \dots]$ is not Noetherian since the ideal $(x_1, x_2, \dots)$ cannot be generated by any finite set (any finite set of generators involves only finitely many of the x_i). Exercise 33(d) in Section 7.4 shows that the ring of continuous real valued functions on $[0, 1]$ is not Noetherian.

A Noetherian ring may have arbitrarily long ascending chains of ideals and may have infinitely long descending chains of ideals. For example, $\mathbb{Z}$ has the infinite descending chain

$$(2) \supset (4) \supset (8) \supset \cdots$$

i.e., a Noetherian ring need not satisfy the *descending chain condition on ideals (D.C.C.)*. We shall see, however, that a commutative ring satisfying D.C.C. on ideals necessarily also satisfies A.C.C., i.e., is Noetherian; such rings are called *Artinian* and are studied in Chapter 16.

The following theorem and its corollary, which we record here for completeness, were proved in Section 9.6 (Theorem 21 and Corollary 22, respectively).

Theorem 3. *(Hilbert's Basis Theorem)* If R is a Noetherian ring then so is the polynomial ring $R[x]$.

Note that Hilbert's Basis Theorem shows how larger Noetherian rings may be built from existing ones in a manner analogous to Theorem 7 of Section 9.3 (which proved that if R is a U.F.D., then so is $R[x]$).

Corollary 4. The polynomial ring $k[x_1, x_2, \dots, x_n]$ with coefficients from a field k is a Noetherian ring.

Let k be a field. Recall that a ring R is a *k-algebra* if k is contained in the center of R and the identity of k is the identity of R.

Definition.

(1) The ring R is a *finitely generated k-algebra* if R is generated as a ring by k together with some finite set $r_1, r_2, \dots, r_n$ of elements of R.

(2) Let R and S be k-algebras. A map $\psi : R \to S$ is a *k-algebra homomorphism* if ψ is a ring homomorphism that is the identity on k.

If R is a k-algebra then R is both a ring and a vector space over k, and it is important to distinguish the sense in which elements of R are generators for R. For example, the polynomial ring $k[x_1, \dots, x_n]$ in a finite number of variables over k is a finitely generated k-algebra since $x_1, \dots, x_n$ are ring generators, but for $n > 0$ this ring is an *infinite* dimensional vector space over k.

Corollary 5. The ring R is a finitely generated k-algebra if and only if there is some surjective k-algebra homomorphism

$$\varphi : k[x_1, x_2, \dots, x_n] \to R$$

from the polynomial ring in a finite number of variables onto R that is the identity map on k. Any finitely generated k-algebra is therefore Noetherian.

Proof: If R is generated as a k-algebra by $r_1, \dots, r_n$, then we may define the map $\varphi : k[x_1, \dots, x_n] \to R$ by $\varphi(x_i) = r_i$ for all i and $\varphi.(a) = a$ for all $a \in k$. Then φ extends uniquely to a surjective ring homomorphism. Conversely, given a surjective homomorphism φ, the images of $x_1, \dots, x_n$ under φ then generate R as a k-algebra, proving that R is finitely generated. Since $k[x_1, \dots, x_n]$ is Noetherian by the previous corollary, any finitely generated k-algebra is therefore the quotient of a Noetherian ring, hence also Noetherian by Proposition 1.

Example

Suppose the k-algebra R is finite dimensional as a vector space over k, for example when $R = k[x]/(f(x))$, where f is any nonzero polynomial in $k[x]$. Then in particular R is a finitely generated k-algebra since a vector space basis also generates R as a ring. In this case since ideals are also k-subspaces any ascending or descending chain of ideals has at most $\dim_k R + 1$ distinct terms, hence R satisfies both A.C.C. and D.C.C. on ideals.

The basic idea behind "algebraic geometry" is to equate geometric questions with algebraic questions involving ideals in rings such as $k[x_1, \dots, x_n]$. The Noetherian nature of these rings reduces many questions to consideration of finitely many algebraic equations (and this was in turn one of the main original motivations for Hilbert's Basis Theorem). We first consider the principal geometric object, the notion of an "algebraic set" of points.

Affine Algebraic Sets

Recall that the set $\mathbb{A}^n$ of n-tuples of elements of the field k is called *affine n-space over k* (cf. Section 10.1). If $x_1, x_2, \dots, x_n$ are independent variables over k, then the polynomials f in $k[x_1, x_2, \dots, x_n]$ can be viewed as k-valued functions $f : \mathbb{A}^n \to k$ on $\mathbb{A}^n$ by evaluating f at the points in $\mathbb{A}^n$:

$$f : (a_1, a_2, \dots, a_n) \mapsto f(a_1, a_2, \dots, a_n) \in k.$$

This gives a ring of k-valued functions on $\mathbb{A}^n$, denoted by $k[\mathbb{A}^n]$ and called the *coordinate ring of* $\mathbb{A}^n$. For instance, when $k = \mathbb{R}$ and $n = 2$, the coordinate ring of Euclidean 2-space $\mathbb{R}^2$ is denoted by $\mathbb{R}[\mathbb{A}^2]$ and is the ring of polynomials in two variables, say x and y, acting as real valued functions on $\mathbb{R}^2$ (the usual "coordinate functions").

Each subset S of functions in the coordinate ring $k[\mathbb{A}^n]$ determines a subset $\mathcal{Z}(S)$ of affine space, namely the set of points where all functions in S are simultaneously zero:

$$\mathcal{Z}(S) = \{(a_1, a_2, \dots, a_n) \in \mathbb{A}^n \mid f(a_1, a_2, \dots, a_n) = 0 \text{ for all } f \in S\},$$

where $\mathcal{Z}(\emptyset) = \mathbb{A}^n$.

Definition. A subset V of $\mathbb{A}^n$ is called an *affine algebraic set* (or just an algebraic set) if V is the set of common zeros of some set S of polynomials, i.e., if $V = \mathcal{Z}(S)$ for some $S \subseteq k[\mathbb{A}^n]$. In this case $V = \mathcal{Z}(S)$ is called the *locus of* S in $\mathbb{A}^n$.

If $S = \{f\}$ or $\{f_1, \dots, f_m\}$ we shall simply write $\mathcal{Z}(f)$ or $\mathcal{Z}(f_1, \dots, f_m)$ for $\mathcal{Z}(S)$ and call it the locus of f or $f_1, \dots, f_m$, respectively. Note that the locus of a single polynomial of the form $f - g$ is the same as the solutions in affine n-space of the equation $f = g$, so affine algebraic sets are the solution sets to systems of polynomial equations, and as a result occur frequently in mathematics.

Examples

(1) If $n = 1$ then the locus of a single polynomial $f \in k[x]$ is the set of roots of f in k. The algebraic sets in $\mathbb{A}^1$ are $\emptyset$, any finite set, and k (cf. the exercises).

(2) The one point subsets of $\mathbb{A}^n$ for any n are affine algebraic since $\{(a_1, a_2, \dots, a_n)\}$ is $\mathcal{Z}(x_1 - a_1,\ x_2 - a_2,\ \dots,\ x_n - a_n)$. More generally, any finite subset of $\mathbb{A}^n$ is an affine algebraic set.

(3) One may define lines, planes, etc. in $\mathbb{A}^n$ — these are *linear algebraic sets*, the loci of sets of linear (degree 1) polynomials of $k[x_1, \dots, x_n]$. For example, a line in $\mathbb{A}^2$ is defined by an equation $ax + by = c$ (which is the locus of the polynomial $f(x, y) = ax + by - c \in k[x, y]$). A line in $\mathbb{A}^3$ is the locus of two linear polynomials of $k[x, y, z]$ that are not multiples of each other. In particular, the coordinate axes, coordinate planes, etc. in $\mathbb{A}^n$ are all affine algebraic sets. For instance, the x-axis in $\mathbb{A}^3$ is the zero set $\mathcal{Z}(y, z)$ and the x,y plane is the zero set $\mathcal{Z}(z)$.

(4) In general the algebraic set $\mathcal{Z}(f)$ of a nonconstant polynomial f is called a *hypersurface* in $\mathbb{A}^n$. Conic sections are familiar algebraic sets in the Euclidean plane $\mathbb{R}^2$. For example, the locus of $y - x^2$ is the parabola $y = x^2$, the locus of $x^2 + y^2 - 1$ is the unit circle, and $\mathcal{Z}(xy - 1)$ is the hyperbola $y = 1/x$. The x- and y-axes are the algebraic sets $\mathcal{Z}(y)$ and $\mathcal{Z}(x)$ respectively. Likewise, quadric surfaces such as the ellipsoid defined by the equation $x^2 + \dfrac{y^2}{4} + \dfrac{z^2}{9} = 1$ are affine algebraic sets in $\mathbb{R}^3$.

We leave as exercises the straightforward verification of the following properties of affine algebraic sets. Let S and T be subsets of $k[\mathbb{A}^n]$.

(1) If $S \subseteq T$ then $\mathcal{Z}(T) \subseteq \mathcal{Z}(S)$ (i.e., $\mathcal{Z}$ is inclusion reversing or *contravariant*).

(2) $\mathcal{Z}(S) = \mathcal{Z}(I)$, where $I = (S)$ is the ideal in $k[\mathbb{A}^n]$ generated by the subset S.

(3) The intersection of two affine algebraic sets is again an affine algebraic set, in fact $\mathcal{Z}(S) \cap \mathcal{Z}(T) = \mathcal{Z}(S \cup T)$. More generally an arbitrary intersection of affine algebraic sets is an algebraic set: if $\{S_j\}$ is any collection of subsets of $k[\mathbb{A}^n]$, then

$$\cap \mathcal{Z}(S_j) = \mathcal{Z}(\cup S_j).$$

(4) The union of two affine algebraic sets is again an affine algebraic set, in fact $\mathcal{Z}(I) \cup \mathcal{Z}(J) = \mathcal{Z}(IJ)$, where I and J are ideals and IJ is their product.

(5) $\mathcal{Z}(0) = \mathbb{A}^n$ and $\mathcal{Z}(1) = \emptyset$ (here 0 and 1 denote constant functions).

By (2), every affine algebraic set is the algebraic set corresponding to an *ideal* of the coordinate ring. Thus we may consider

$$\mathcal{Z} : \{\text{ideals of } k[\mathbb{A}^n]\} \to \{\text{affine algebraic sets in } \mathbb{A}^n\}.$$

Since every ideal I in the Noetherian ring $k[x_1, x_2, \dots, x_n]$ is finitely generated, say $I = (f_1, f_2, \dots, f_q)$, it follows from (3) that $\mathcal{Z}(I) = \mathcal{Z}(f_1) \cap \mathcal{Z}(f_2) \cap \cdots \cap \mathcal{Z}(f_q)$, i.e., *each affine algebraic set is the intersection of a finite number of hypersurfaces in* $\mathbb{A}^n$. Note that this "geometric" property in affine n-space is a consequence of an "algebraic" property of the corresponding coordinate ring (namely, Hilbert's Basis Theorem).

If V is an algebraic set in affine n-space, then there may be many ideals I such that $V = \mathcal{Z}(I)$. For example, in affine 2-space over $\mathbb{R}$ the y-axis is the locus of the ideal (x) of $\mathbb{R}[x, y]$, and also the locus of (x^2), (x^3), etc. More generally, the zeros of any polynomial are the same as the zeros of all its positive powers, and it follows that $\mathcal{Z}(I) = \mathcal{Z}(I^k)$ for all $k \geq 1$. We shall study the relationship between ideals that determine the same affine algebraic set in the next section when we discuss radicals of ideals.

While the ideal whose locus determines a particular algebraic set V is not unique, there is a unique largest ideal that determines V, given by the set of *all* polynomials that vanish on V. In general, for any subset A of $\mathbb{A}^n$ define

$$\mathcal{I}(A) = \{f \in k[x_1, \dots, x_n] \mid f(a_1, a_2, \dots, a_n) = 0 \text{ for all } (a_1, a_2, \dots, a_n) \in A\}.$$

It is immediate that $\mathcal{I}(A)$ is an *ideal*, and is the unique largest ideal of functions that are identically zero on A. This defines a correspondence

$$\mathcal{I} : \{\text{subsets in } \mathbb{A}^n\} \to \{\text{ideals of } k[\mathbb{A}^n]\}.$$

Examples

(1) In the Euclidean plane, $\mathcal{I}$(the x-axis) is the ideal generated by y in the coordinate ring $\mathbb{R}[x, y]$.

(2) Over any field k, the ideal of functions vanishing at $(a_1, a_2, \dots, a_n) \in \mathbb{A}^n$ is a maximal ideal since it is the kernel of the surjective ring homomorphism from $k[x_1, \dots, x_n]$ to the field k given by evaluation at $(a_1, a_2, \dots, a_n)$. It follows that

$$\mathcal{I}((a_1, a_2, \dots, a_n)) = (x_1 - a_1,\ x_2 - a_2,\ \dots,\ x_n - a_n).$$

(3) Let $V = \mathcal{Z}(x^3 - y^2)$ in $\mathbb{A}^2$. If $(a, b) \in \mathbb{A}^2$ is an element of V then $a^3 = b^2$. If $a \neq 0$, then also $b \neq 0$ and we can write $a = (b/a)^2$, $b = (b/a)^3$. It follows that V is the set $\{(a^2, a^3) \mid a \in k\}$. For any polynomial $f(x, y) \in k[x, y]$ we can write $f(x, y) = f_0(x) + f_1(x)y + (x^3 - y^2)g(x, y)$. For $f(x, y) \in \mathcal{I}(V)$, i.e., $f(a^2, a^3) = 0$ for all $a \in k$, it follows that $f_0(a^2) + f_1(a^2)a^3 = 0$ for all $a \in k$. If $f_0(x) = a_r x^r + \cdots + a_0$ and $f_1(x) = b_s x^s + \cdots + b_0$ then

$$f_0(x^2) + x^3 f_1(x^2) = (a_r x^{2r} + \cdots + a_0) + (b_s x^{2s+3} + \cdots + b_0 x^3)$$

and this polynomial is 0 for every $a \in k$. If k is infinite, this polynomial has infinitely many zeros, which can happen only if all of the coefficients are zero. The coefficients of the terms of even degree are the coefficients of $f_0(x)$ and the coefficients of the terms of odd degree are the coefficients of $f_1(x)$, so it follows that $f_0(x)$ and $f_1(x)$ are both 0. It follows that $f(x, y) = (x^3 - y^2)g(x, y)$, and so

$$\mathcal{I}(V) = (x^3 - y^2) \subset k[x, y].$$

If k is finite, however, there may be elements in $\mathcal{I}(V)$ not lying in the ideal $(x^3 - y^2)$. For example, if $k = \mathbb{F}_2$, then V is simply the set $\{(0, 0), (1, 1)\}$ and so $\mathcal{I}(V)$ contains the polynomial $x(x - 1)$ (cf. Exercise 15).

The following properties of the map $\mathcal{I}$ are very easy exercises. Let A and B be subsets of $\mathbb{A}^n$.

(6) If $A \subseteq B$ then $\mathcal{I}(B) \subseteq \mathcal{I}(A)$ (i.e., $\mathcal{I}$ is also *contravariant*).
(7) $\mathcal{I}(A \cup B) = \mathcal{I}(A) \cap \mathcal{I}(B)$.
(8) $\mathcal{I}(\emptyset) = k[x_1, \dots, x_n]$ and, if k is infinite, $\mathcal{I}(\mathbb{A}^n) = 0$.

Moreover, there are easily verified relations between the maps $\mathcal{Z}$ and $\mathcal{I}$:

(9) If A is any subset of $\mathbb{A}^n$ then $A \subseteq \mathcal{Z}(\mathcal{I}(A))$, and if I is any ideal then $I \subseteq \mathcal{I}(\mathcal{Z}(I))$.
(10) If $V = \mathcal{Z}(I)$ is an affine algebraic set then $V = \mathcal{Z}(\mathcal{I}(V))$, and if $I = \mathcal{I}(A)$ then $\mathcal{I}(\mathcal{Z}(I)) = I$, i.e., $\mathcal{Z}(\mathcal{I}(\mathcal{Z}(I))) = \mathcal{Z}(I)$ and $\mathcal{I}(\mathcal{Z}(\mathcal{I}(A))) = \mathcal{I}(A)$.

The last relation shows that the maps $\mathcal{Z}$ and $\mathcal{I}$ act as inverses of each other provided one restricts to the collection of affine algebraic sets $V = \mathcal{Z}(I)$ in $\mathbb{A}^n$ and to the set of ideals in $k[\mathbb{A}^n]$ of the form $\mathcal{I}(V)$. In the case where the field k is algebraically closed we shall (in the following two sections) characterize those ideals I that are of the form $\mathcal{I}(V)$ for some affine algebraic set V in terms of purely ring-theoretic properties of the ideal I (this is the famous "Zeros Theorem" of Hilbert, cf. Theorem 32).

Definition. If $V \subseteq \mathbb{A}^n$ is an affine algebraic set the quotient ring $k[\mathbb{A}^n]/\mathcal{I}(V)$ is called the *coordinate ring of* V, and is denoted by $k[V]$.

Note that for $V = \mathbb{A}^n$ and k infinite we have $\mathcal{I}(V) = 0$, so this definition extends the previous terminology. The polynomials in $k[\mathbb{A}^n]$ define k-valued functions on V simply by restricting these functions on $\mathbb{A}^n$ to the subset V. Two such polynomial functions f and g define the *same* function on V if and only if $f - g$ is identically 0 on V, which is to say that $f - g \in \mathcal{I}(V)$. Hence the cosets $\overline{f} = f + \mathcal{I}(V)$ giving the elements of the quotient $k[V]$ are precisely the restrictions to V of ordinary polynomial functions f from $\mathbb{A}^n$ to k (which helps to explain the notation $k[V]$). If x_i denotes the i^{th} coordinate function on $\mathbb{A}^n$ (projecting an n-tuple onto its i^{th} component), then the restriction $\overline{x_i}$ of x_i to V (which also just gives the i^{th} component of the elements in V viewed as a subset of $\mathbb{A}^n$) is an element of $k[V]$, and $k[V]$ is finitely generated as a k-algebra by $\overline{x_1}, \dots, \overline{x_n}$ (although this need not be a minimal generating set).

Example

If $V = \mathcal{Z}(xy - 1)$ is the hyperbola $y = 1/x$ in $\mathbb{R}^2$, then $\mathbb{R}[V] = \mathbb{R}[x, y]/(xy - 1)$. The polynomials $f(x, y) = x$ (the x-coordinate function) and $g(x, y) = x + (xy - 1)$, which are different functions on $\mathbb{R}^2$, define the same function on the subset V. On the point $(1/2, 2) \in V$, for example, both give the value $1/2$. In the quotient ring $\mathbb{R}[V]$ we have $\overline{x}\,\overline{y} = 1$, so $\mathbb{R}[V] \cong \mathbb{R}[x, 1/x]$. For any function $\overline{f} \in \mathbb{R}[V]$ and any $(a, b) \in V$ we have $\overline{f}(a, b) = f(a, 1/a)$ for any polynomial $f \in k[x, y]$ mapping to $\overline{f}$ in the quotient.

Suppose now that $V \subseteq \mathbb{A}^n$ and $W \subseteq \mathbb{A}^m$ are two affine algebraic sets. Since V and W are defined by the vanishing of polynomials, the most natural algebraic maps between V and W are those defined by polynomials:

Definition. A map $\varphi : V \to W$ is called a *morphism* (or *polynomial map* or *regular map*) of algebraic sets if there are polynomials $\varphi_1, \dots, \varphi_m \in k[x_1, x_2, \dots, x_n]$ such that

$$\varphi((a_1, \dots, a_n)) = (\varphi_1(a_1, \dots, a_n), \dots, \varphi_m(a_1, \dots, a_n))$$

for all $(a_1, \dots, a_n) \in V$. The map $\varphi : V \to W$ is an *isomorphism* of algebraic sets if there is a morphism $\psi : W \to V$ with $\varphi \circ \psi = 1_W$ and $\psi \circ \varphi = 1_V$.

Note that in general $\varphi_1, \varphi_2, \dots, \varphi_m$ are not uniquely defined. For example, both $f = x$ and $g = x + (xy - 1)$ in the example above define the same morphism from $V = \mathcal{Z}(xy - 1)$ to $W = \mathbb{A}^1$.

Suppose F is a polynomial in $k[x_1, \dots, x_m]$. Then $F \circ \varphi = F(\varphi_1, \varphi_2, \dots, \varphi_m)$ is a polynomial in $k[x_1, \dots, x_n]$ since $\varphi_1, \varphi_2, \dots, \varphi_m$ are polynomials in $x_1, \dots, x_n$. If $F \in \mathcal{I}(W)$, then $F \circ \varphi((a_1, a_2, \dots, a_n)) = 0$ for every $(a_1, a_2, \dots, a_n) \in V$ since $\varphi((a_1, a_2, \dots, a_n)) \in W$. Thus $F \circ \varphi \in \mathcal{I}(V)$. It follows that φ induces a well defined map from the quotient ring $k[x_1, \dots, x_m]/\mathcal{I}(W)$ to the quotient ring $k[x_1, \dots, x_n]/\mathcal{I}(V)$:

$$\begin{aligned} \tilde{\varphi} : k[W] &\to k[V] \\ f &\mapsto f \circ \varphi \end{aligned}$$

where $f \circ \varphi$ is given by $F \circ \varphi + \mathcal{I}(V)$ for any polynomial $F = F(x_1, \dots, x_m)$ with $f = F + \mathcal{I}(W)$. It is easy to check that $\tilde{\varphi}$ is a k-algebra homomorphism (for example, $\tilde{\varphi}(f + g) = (f + g) \circ \varphi = f \circ \varphi + g \circ \varphi = \tilde{\varphi}(f) + \tilde{\varphi}(g)$ shows that $\tilde{\varphi}$ is additive). Note also the contravariant nature of $\tilde{\varphi}$: the morphism from V to W induces a k-algebra homomorphism from $k[W]$ to $k[V]$.

Suppose conversely that Φ is any k-algebra homomorphism from the coordinate ring $k[W] = k[x_1, \dots, x_m]/\mathcal{I}(W)$ to $k[V] = k[x_1, \dots, x_n]/\mathcal{I}(V)$. Let F_i be a representative in $k[x_1, \dots, x_n]$ for the image under Φ of $\bar{x}_i \in k[W]$ (i.e., $\Phi(x_i \bmod \mathcal{I}(W))$ is $F_i \bmod \mathcal{I}(V)$). Then $\varphi = (F_1, \dots, F_m)$ defines a polynomial map from $\mathbb{A}^n$ to $\mathbb{A}^m$, and in fact φ is a morphism from V to W. To see this it suffices to check that φ maps a point of V to a point of W since by definition φ is already defined by polynomials. If $g \in \mathcal{I}(W) \subset k[x_1, \dots, x_m]$, then in $k[W]$ we have

$$g(x_1 + \mathcal{I}(W), \dots, x_m + \mathcal{I}(W)) = g(x_1, \dots, x_m) + \mathcal{I}(W) = \mathcal{I}(W) = 0 \in k[W],$$

and so

$$\Phi(g(x_1 + \mathcal{I}(W), \dots, x_m + \mathcal{I}(W))) = 0 \in k[V].$$

Since Φ is a k-algebra homomorphism, it follows that

$$g(\Phi(x_1 + \mathcal{I}(W)), \dots, \Phi(x_m + \mathcal{I}(W)) = 0 \in k[V].$$

By definition, $\Phi(x_i + \mathcal{I}(W)) = F_i \bmod \mathcal{I}(V)$, so

$$g(F_1 \bmod \mathcal{I}(V), \dots, F_m \bmod \mathcal{I}(V)) = 0 \in k[V],$$

i.e.,

$$g(F_1, \dots, F_m) \in \mathcal{I}(V).$$

It follows that $g(F_1(a_1, \dots, a_n), \dots, F_m(a_1, \dots, a_n)) = 0$ for every $(a_1, \dots, a_n)$ in V. This shows that if $(a_1, \dots, a_n) \in V$, then every polynomial in $\mathcal{I}(W)$ vanishes on $\varphi(a_1, \ldots, a_n)$. By property (10) of the maps $\mathcal{Z}$ and $\mathcal{I}$ above, this means that $\varphi(a_1, \ldots, a_n) \in \mathcal{Z}(\mathcal{I}(W)) = W$, which proves that φ maps a point in V to a point in W. It follows that $\varphi = (F_1, \ldots, F_m)$ is a morphism from V to W. Since the F_i are well defined modulo $\mathcal{I}(V)$, this morphism from V to W does not depend on the choice of the F_i. Furthermore, the morphism φ induces the original k-algebra homomorphism Φ from $k[W]$ to $k[V]$, i.e., $\widetilde{\varphi} = \Phi$, since both homomorphisms take the value $F_i + \mathcal{I}(V)$ on $x_i + \mathcal{I}(W) \in k[W]$. This proves the first two statements in the following theorem.

Theorem 6. Let $V \subseteq \mathbb{A}^n$ and $W \subseteq \mathbb{A}^m$ be affine algebraic sets. Then there is a bijective correspondence

$$\left\{ \begin{array}{c} \text{morphisms from } V \text{ to } W \\ \text{as algebraic sets} \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} k\text{-algebra homomorphisms} \\ \text{from } k[W] \text{ to } k[V] \end{array} \right\}.$$

More precisely,

(1) Every morphism $\varphi : V \to W$ induces an associated k-algebra homomorphism $\widetilde{\varphi} : k[W] \to k[V]$ defined by $\widetilde{\varphi}(f) = f \circ \varphi$.

(2) Every k-algebra homomorphism $\Phi : k[W] \to k[V]$ is induced by a unique morphism $\varphi : V \to W$, i.e., $\Phi = \widetilde{\varphi}$.

(3) If $\varphi : V \to W$ and $\psi : W \to U$ are morphisms of affine algebraic sets, then $\widetilde{\psi \circ \varphi} = \widetilde{\varphi} \circ \widetilde{\psi} : k[U] \to k[V]$.

(4) The morphism $\varphi : V \to W$ is an isomorphism if and only if $\widetilde{\varphi} : k[W] \to k[V]$ is a k-algebra isomorphism.

Proof: The proof of (3) is left as an exercise and (4) is then immediate.

Example

For any infinite field k let $V = \mathbb{A}^1$ and let $W = \mathcal{Z}(x^3 - y^2) = \{(a^2, a^3) \mid a \in k\}$. The map $\varphi : V \to W$ defined by $\varphi(a) = (a^2, a^3)$ is a morphism from V to W. Note that φ is a bijection. The coordinate rings are $k[V] = k[x]$ and $k[W] = k[x, y]/(x^3 - y^2)$ (by the computations in a previous example — it is at this point we need k to be infinite) and the associated k-algebra homomorphism of coordinate rings is determined by

$$\begin{aligned} \widetilde{\varphi} : k[W] &\longrightarrow k[V] \\ x &\mapsto x^2 \\ y &\mapsto x^3. \end{aligned}$$

The image of $\widetilde{\varphi}$ is the subalgebra $k[x^2, x^3] = k + x^2 k[x]$ of $k[x]$, so in particular $\widetilde{\varphi}$ is not surjective. Hence $\widetilde{\varphi}$ is not an isomorphism of coordinate rings, and it follows that φ is not an isomorphism of algebraic sets, even though the morphism φ is a bijective map. The inverse map is given by $\psi(0, 0) = 0$ and $\psi(a, b) = b/a$ for $b \neq 0$, and this cannot be achieved by a polynomial map.

The bijection in Theorem 6 gives a translation from maps between two geometrically defined algebraic sets V and W into algebraic maps between their coordinate rings. It also allows us to define a morphism intrinsically in terms of V and W without explicit reference to the ambient affine spaces containing them:

Corollary 7. Suppose $\varphi : V \to W$ is a map of affine algebraic sets. Then φ is a morphism if and only if for every $f \in k[W]$ the composite map $f \circ \varphi$ is an element of $k[V]$ (as a k-valued function on V). When φ is a morphism, $\varphi(v) = w$ with $v \in V$ and $w \in W$ if and only if $\widetilde{\varphi}^{-1}(\mathcal{I}(\{v\})) = \mathcal{I}(\{w\})$.

Proof: We first prove that if φ is any map from V to W such that $\widetilde{\varphi}$ is a k-algebra homomorphism then $\varphi(v) = w$ if and only if $\widetilde{\varphi}^{-1}(\mathcal{I}(\{v\})) = \mathcal{I}(\{w\})$, which will in particular establish the second statement. Note that $\varphi(v) = w$ if and only if every polynomial f vanishing at w vanishes at $\varphi(v)$ (by property (10) above: $\{w\} = \mathcal{Z}(\mathcal{I}(\{w\}))$). Since f vanishes at $\varphi(v)$ if and only if $\widetilde{\varphi}(f)$ vanishes at v, this is equivalent to the statement that $\widetilde{\varphi}(f) \in \mathcal{I}(\{v\})$ for every $f \in \mathcal{I}(\{w\})$, i.e., $\widetilde{\varphi}(\mathcal{I}(\{w\})) \subseteq \mathcal{I}(\{v\})$, or $\mathcal{I}(\{w\}) \subseteq \widetilde{\varphi}^{-1}(\mathcal{I}(\{v\}))$. Since both $\mathcal{I}(\{w\})$ and $\mathcal{I}(\{v\})$ are maximal ideals, this is equivalent to $\widetilde{\varphi}^{-1}(\mathcal{I}(\{v\})) = \mathcal{I}(\{w\})$.

We now prove the first statement. If φ is a morphism, then $f \circ \varphi \in k[V]$ for every $f \in k[W]$. For the converse, observe first that composition with any map $\varphi : V \to W$ defines a k-algebra homomorphism $\widetilde{\varphi}$ from the k-algebra of k-valued functions on W to the k-algebra of k-valued functions on V (this is immediate from the pointwise definition of the addition and multiplication of functions). If $f \circ \varphi \in k[V]$ for every $f \in k[W]$, then $\widetilde{\varphi}$ is a k-algebra homomorphism from $k[W]$ to $k[V]$, so by the proposition, $\widetilde{\varphi} = \widetilde{\Phi}$ for a unique morphism $\Phi : V \to W$. Also, since $\widetilde{\varphi}$ is a k-algebra homomorphism from $k[W]$ to $k[V]$ it follows by what we have already shown that $\varphi(v) = w$ if and only if $\widetilde{\varphi}^{-1}(\mathcal{I}(\{v\})) = \mathcal{I}(\{w\})$. Because $\widetilde{\varphi} = \widetilde{\Phi}$, this is equivalent to $\widetilde{\Phi}^{-1}(\mathcal{I}(\{v\})) = \mathcal{I}(\{w\})$, and so $\Phi(v) = w$. Hence φ and Φ define the same map on V and so φ is a morphism, completing the proof.

Corollary 7 and the last part of Theorem 6 show that the isomorphism type of the coordinate ring of V (as a k-algebra) does not depend on the embedding of V in a particular affine n-space.

Computations in Affine Algebraic Sets and k-algebras

The theory of Gröbner bases developed in Section 9.6 is very useful in computations involving affine algebraic sets, for example in computing in the coordinate rings $k[\mathbb{A}^n]/\mathcal{I}(V)$. When $n > 1$ it can be difficult to describe the elements in this quotient ring explicitly. By Theorem 23 in Section 9.6, each polynomial f in $k[\mathbb{A}^n]$ has a unique remainder after general polynomial division by the elements in a Gröbner basis for $\mathcal{I}(V)$, and this remainder therefore serves as a unique representative for the coset $\bar{f}$ of f in the quotient $k[\mathbb{A}^n]/\mathcal{I}(V)$.

Examples

(1) In the example $W = \mathcal{Z}(x^3 - y^2)$ above, we showed $I = \mathcal{I}(W) = (x^3 - y^2)$ for any infinite field k and so $k[W] = k[x, y]/(x^3 - y^2)$. Here $x^3 - y^2$ gives a Gröbner basis for I with respect to the lexicographic monomial ordering with $y > x$, so every polynomial $f = f(x, y)$ can be written uniquely in the form $f(x, y) = f_0(x) + f_1(x)y + f_I$ with $f_0(x), f_1(x) \in k[x]$ and $f_I \in I$. Then $f_0(x) + f_1(x)y$ gives a unique representative for $\bar{f}$ in $k[W]$. With respect to the lexicographic monomial ordering with $x > y$,

$x^3 - y^2$ is again a Gröbner basis for I, but now the remainder representing $\bar{f}$ in $k[W]$ is of the form $h_0(y) + h_1(y)x + h_2(y)x^2$.

(2) Let $V = \mathcal{Z}(xz+y^2+z^2, xy-xz+yz-2z^2) \subset \mathbb{C}^3$ and $W = \mathcal{Z}(u^3-uv^2+v^3) \subset \mathbb{C}^2$. We shall show later that $I = \mathcal{I}(V) = (xz+y^2+z^2, xy-xz+yz-2z^2) \subset \mathbb{C}[x, y, z]$ and $J = \mathcal{I}(W) = (u^3 - uv^2 + v^3) \subset \mathbb{C}[u, v]$. In this case $u^3 - uv^2 + v^3$ gives a Gröbner basis for J for the lexicographic monomial ordering with $u > v$ similar to the previous example. The situation for I is more complicated. With respect to the lexicographic monomial ordering with $x > y > z$ the reduced Gröbner basis for I is given by

$$g_1 = xy + y^2 + yz - z^2, \qquad g_2 = xz + y^2 + z^2, \qquad g_3 = y^3 - y^2z + z^3.$$

Unique representatives for $\mathbb{C}[V] = \mathbb{C}[x, y, z]/(x^2 + xz + y^2, 2x^2 - xy + xz - yz)$ are given by the remainders after general polynomial division by $\{g_1, g_2, g_3\}$.

We saw already in Section 9.6 that Gröbner bases and elimination theory can be used in the explicit computation of affine algebraic sets $\mathcal{Z}(S)$, or, equivalently, in explicitly solving systems of algebraic equations. The same theory can be used to determine explicitly a set of generators for the image and kernel of a k-algebra homomorphism

$$\Phi : k[y_1, \ldots, y_m]/J \longrightarrow k[x_1, \ldots, x_n]/I$$

where I and J are ideals. In the particular case when $I = \mathcal{I}(V)$ and $J = \mathcal{I}(W)$ are the ideals associated to affine algebraic sets $V \subseteq \mathbb{A}^n$ and $W \subseteq \mathbb{A}^m$ then by Theorem 6, the k-algebra homomorphism Φ corresponds to a morphism from V to W, and we shall apply the results here to affine algebraic sets in Section 3.

For $1 \le i \le m$, let $\varphi_i \in k[x_1, \ldots, x_n]$ be any polynomial representing the coset $\Phi(\bar{y}_i)$, where as usual we use a bar to denote the coset of an element in a quotient. The polynomials $\varphi_1, \ldots, \varphi_n$ are unique up to elements of I. Then the image of a coset $f(y_1, \ldots, y_m) + J$ under Φ is the coset $f(\varphi_1, \ldots, \varphi_m) + I$. Given any $\varphi_1, \ldots, \varphi_n$, the map sending y_i to φ_i induces a k-algebra homomorphism Φ if and only if $f(y_1, \ldots, y_m) \in I$ for every $f \in J$, a condition which can be checked on a set of generators for J.

Proposition 8. With notation as above, let $R = k[y_1, \ldots, y_m, x_1, \ldots, x_n]$ and let $\mathcal{A}$ be the ideal generated by $y_1 - \varphi_1, \ldots, y_m - \varphi_m$ together with generators for I. Let G be the reduced Gröbner basis of $\mathcal{A}$ with respect to the lexicographic monomial ordering $x_1 > \cdots > x_n > y_1 > \cdots > y_m$. Then

(a) The kernel of Φ is $\mathcal{A} \cap k[y_1, \ldots, y_m]$ modulo J. The elements of G in $k[y_1, \ldots, y_m]$ (taken modulo J) generate $\ker \Phi$.

(b) If $f \in k[x_1, \ldots, x_n]$, then $\bar{f}$ is in the image of Φ if and only if the remainder after general polynomial division of f by the elements in G is an element $h \in k[y_1, \ldots, y_m]$, in which case $\Phi(\bar{h}) = \bar{f}$.

Proof: If we show $\ker \Phi = \mathcal{A} \cap k[y_1, \ldots, y_m]$ modulo J then (a) follows by Proposition 30 in Section 9.6. Suppose first that $f \in \mathcal{A} \cap k[y_1, \ldots, y_m]$. If $f_1, \ldots, f_s$ are generators for I in $k[x_1, \ldots, x_n]$, then

$$f(y_1, \ldots, y_m) = \sum_{i=1}^{n} a_i(y_i - \varphi_i) + \sum_{j=1}^{s} b_i f_i$$

as polynomials in R, where $a_1, \dots, a_n, b_1, \dots, b_s \in R$. Substituting $y_i = \varphi_i$ we see that $f(\varphi_1, \dots, \varphi_m)$ is an element of I. Since $\Phi(\bar{f}) = f(\varphi_1, \dots, \varphi_m)$ modulo I, it follows that f represents a coset in the kernel of Φ. Conversely, suppose $f \in k[y_1, \dots, y_m]$ represents an element in $\ker \Phi$. Then $f(\varphi_1, \dots, \varphi_m) \in I$ (in $k[x_1, \dots, x_n]$) and so also $f(\varphi_1, \dots, \varphi_m) \in \mathcal{A}$ (in R). Since $y_i - \varphi_i \in \mathcal{A}$,

$$f(y_1, \dots, y_m) \equiv f(\varphi_1, \dots, \varphi_m) \equiv 0 \bmod \mathcal{A}$$

so $f \in \mathcal{A} \cap k[y_1, \dots, y_m]$.

For (b), suppose first that $f \in k[x_1, \dots, x_n]$ represents an element in the image of Φ, i.e., $\bar{f} = \Phi(\bar{h})$ for some polynomial $h \in k[y_1, \dots, y_m]$. Then

$$f(x_1, \dots, x_n) - h(\varphi_1, \dots, \varphi_m) \in I$$

as polynomials in $k[x_1, \dots, x_n]$, and so $f(x_1, \dots, x_n) - h(\varphi_1, \dots, \varphi_m) \in \mathcal{A}$ as polynomials in R. As before, since each $y_i - \varphi_i \in \mathcal{A}$ it follows that

$$f(x_1, \dots, x_n) - h(y_1, \dots, y_m) \in \mathcal{A}.$$

Then $f(x_1, \dots, x_n)$ and $h(y_1, \dots, y_m)$ leave the same remainder after general polynomial division by the elements in G. Since $x_1 > \cdots > x_n > y_1 > \cdots > y_m$, the remainder of $h(y_1, \dots, y_m)$ is again a polynomial h_0 only involving $y_1, \dots, y_m$. Note also that $h - h_0 \in \mathcal{A} \cap k[y_1, \dots, y_m]$ so $\bar{h}$ and $\overline{h_0}$ differ by an element in $\ker \Phi$ by (a), so $\Phi(\overline{h_0}) = \Phi(\bar{h}) = \bar{f}$. For the converse, if f leaves the remainder $h \in k[y_1, \dots, y_m]$ after general polynomial division by the elements in G then $f(x_1, \dots, x_n) - h(y_1, \dots, y_m) \in \mathcal{A}$, i.e.,

$$f(x_1, \dots, x_n) - h(y_1, \dots, y_m) = \sum_{i=1}^{n} a_i(y_i - \varphi_i) + \sum_{j=1}^{s} b_i f_i$$

as polynomials in R, where $a_1, \dots, a_n, b_1, \dots, b_s \in R$. Substituting $y_i = \varphi_i$ we obtain

$$f(x_1, \dots, x_n) - h(\varphi_1, \dots, \varphi_m) \in I$$

as polynomials in $x_1, \dots x_n$, and so $\bar{f} = \Phi(\bar{h})$.

It follows in particular from Proposition 8 that Φ will be a surjective homomorphism if and only if for each $i = 1, 2, \dots, n$, dividing x_i by the elements in the Gröbner basis G leaves a remainder h_i in $k[y_1, \dots, y_m]$. In particular, $x_n - h_n$ leaves a remainder of 0. But this means the leading term of some element g_n in G divides the leading term of $x_n - h_n$ and since $x_1 > \cdots > x_n > y_1 > \cdots > y_m$ by the choice of the ordering, the leading term of $x_n - h_n$ is just x_n. It follows that $LT(g_n) = x_n$ and so $g_n = x_n - h_{n,0} \in G$ for some $h_{n,0} \in k[y_1, \dots, y_m]$ (in fact $h_{n,0}$ is the remainder of h_n after division by the elements in G). Next, since $x_{n-1} - h_{n-1}$ leaves a remainder of 0, there is an element g_{n-1} in G whose leading term is x_{n-1}. Since G is a reduced Gröbner basis and $g_n \in G$, the leading term of g_n, i.e., x_n, does not divide any of the terms in g_{n-1} and it follows that $g_{n-1} = x_{n-1} - h_{n-1,0} \in G$ for some $h_{n-1,0} \in k[y_1, \dots, y_m]$. Proceeding in a similar fashion we obtain the following corollary, showing that whether Φ is surjective can be seen immediately from the elements in the reduced Gröbner basis.

Corollary 9. The map Φ is surjective if and only if for each i, $1 \leq i \leq n$, the reduced Gröbner basis G contains a polynomial $x_i - h_i$ where $h_i \in k[y_1, \dots, y_m]$.

Examples

(1) Let $\Phi : \mathbb{Q}[u, v] \to \mathbb{Q}[x]$ be defined by $\Phi(u) = x^2 + x$ and $\Phi(v) = x^3$. The reduced Gröbner basis G for the ideal $\mathcal{A} = (u - x^2 - x, v - x^3)$ with respect to the lexicographic monomial ordering $x > u > v$ is

$$g_1 = x^2 + x - u, \qquad g_3 = vx - x - u^2 + u + 2v,$$

$$g_2 = ux + x - u - v, \qquad g_4 = u^3 - 3uv - v^2 - v.$$

The kernel of Φ is the ideal generated by $G \cap \mathbb{Q}[u, v] = \{g_4\}$. By Corollary 9, we see that Φ is not surjective. The remainder after general polynomial division of x^4 by $\{g_1, g_s, g_3, g_4\}$ is $x + u^2 - u - 2v \notin \mathbb{Q}[u, v]$, so x^4 is not in the image of Φ. The remainder of $x^5 + x$ is $-u^2 + uv + u + 2v \in \mathbb{Q}[u, v]$ so $x^5 + x = \Phi(-u^2 + uv + u + 2v)$ is in the image of Φ, as a quick check will confirm.

(2) Let $V = \mathcal{Z}(I) \subset \mathbb{C}^3$ and $W = \mathcal{Z}(J) \subset \mathbb{C}^2$ where $I = (xz + y^2 + z^2, xy - xz + yz - 2z^2)$ and $J = (u^3 - uv^2 + v^3)$ as in Example 2 following Corollary 7. Then the map $\varphi : V \to W$ defined by $\varphi((a, b, c)) = (c, b)$ is a morphism from V to W. To see this, we must check that $(c, b) \in W$ if $(a, b, c) \in V$. Equivalently, by Theorem 6, we must check that the map

$$\widetilde{\varphi} : \mathbb{C}[u, v]/(u^3 - uv^2 + v^3) \longrightarrow \mathbb{C}[x, y, z]/(xz + y^2 + z^2, xy - xz + yz - 2z^2)$$

induced by mapping u to z and v to y is a $\mathbb{C}$-algebra homomorphism. This in turn is equivalent to verifying that $f = z^3 - zy^2 + y^3$ is an element of the ideal I. In this case f is actually an element in the reduced Gröbner basis for I:

$$xy + y^2 + yz - z^2, \qquad xz + y^2 + z^2, \qquad y^3 - y^2 z + z^3,$$

so certainly $f \in I$. (Note that dividing f by the original two generators for I leaves the nonzero remainder f itself, from which it is much less clear that $f \in I$, so it is important to use a Gröbner basis when working in coordinate rings.)

(3) In the previous example, let $\mathcal{A} = (u - z, v - y, xz + y^2 + z^2, xy - xz + yz - 2z^2) \subset \mathbb{C}[u, v, x, y, z]$ as in Proposition 8. With respect to the lexicographic monomial ordering $x > y > z > u > v$ the reduced Gröbner basis G for $\mathcal{A}$ is

$$xu + u^2 + v^2, \quad xv - u^2 + uv + v^2, \quad y - v, \quad z - u, \quad u^3 - uv^2 + v^3.$$

By Proposition 8, we see that $\ker \widetilde{\varphi}$ is generated by $u^3 - uv^2 + v^3 \equiv 0 \bmod J$, so $\widetilde{\varphi}$ is injective. Since there is no element of the form $x - h(u, v)$ in G, $\widetilde{\varphi}$ is not surjective (in fact x is not in the image).

As a final example, we use the determination of the kernel of k-algebra homomorphisms to compute minimal polynomials of elements in simple algebraic field extensions.

Proposition 10. Suppose α is a root of the irreducible polynomial $p(x) \in k[x]$ and $\beta \in k(\alpha)$, say $\beta = f(\alpha)$ for the polynomial $f \in k[x]$. Let G be the reduced Gröbner basis for the ideal $(p, y - f)$ in $k[x, y]$ for the lexicographic monomial ordering $x > y$. Then the minimal polynomial of β over k is the monic polynomial in $G \cap k[y]$.

Proof: The kernel of the k-algebra homomorphism $k[y] \to k[x]/(p) \cong k(\alpha)$ defined by mapping y first to f and then to β is the principal ideal generated by the minimal polynomial of β in $k[y]$, and the result follows by Proposition 8.

Example

Take $k = \mathbb{Q}$, and let $\beta = 1+\sqrt[3]{2}+3\sqrt[3]{4} \in \mathbb{Q}(\sqrt[3]{2})$. Then the ideal $(x^3-2, y-(1+x+3x^2))$ in $\mathbb{Q}[x, y]$ has reduced Gröbner basis $\{53x - 3y^2 + 7y + 32, y^3 - 3y^2 - 15y - 93\}$ for the lexicographic monomial ordering $x > y$, so the minimal polynomial for β is $y^3 - 3y^2 - 15y - 93$.

EXERCISES

Let R be a commutative ring with $1 \neq 0$ and let k be a field.

1. Prove the converse to Hilbert's Basis Theorem: if the polynomial ring $R[x]$ is Noetherian, then R is Noetherian.
2. Show that each of the following rings are not Noetherian by exhibiting an explicit infinite increasing chain of ideals:
   (a) the ring of continuous real valued functions on $[0, 1]$,
   (b) the ring of all functions from any infinite set X to $\mathbb{Z}/2\mathbb{Z}$.
3. Prove that the field $k(x)$ of rational functions over k in the variable x is not a finitely generated k-algebra. (Recall that $k(x)$ is the field of fractions of the polynomial ring $k[x]$. Note that $k(x)$ *is* a finitely generated *field extension* over k.)
4. Prove that if R is Noetherian, then so is the ring $R[[x]]$ of formal power series in the variable x with coefficients from R (cf. Exercise 3, Section 7.2). [Mimic the proof of Hilbert's Basis Theorem.]
5. (*Fitting's Lemma*) Suppose M is a Noetherian R-module and $\varphi : M \to M$ is an R-module endomorphism of M. Prove that $\ker(\varphi^n) \cap \text{image}(\varphi^n) = 0$ for n sufficiently large. Show that if φ is surjective, then φ is an isomorphism. [Observe that $\ker(\varphi) \subseteq \ker(\varphi^2) \subseteq \dots$.]
6. Suppose that $0 \longrightarrow M' \longrightarrow M \longrightarrow M'' \longrightarrow 0$ is an exact sequence of R-modules. Prove that M is a Noetherian R-module if and only if M' and M'' are Noetherian R-modules.
7. Prove that submodules, quotient modules, and finite direct sums of Noetherian R-modules are again Noetherian R-modules.
8. If R is a Noetherian ring, prove that M is a Noetherian R-module if and only if M is a finitely generated R-module. (Thus any submodule of a finitely generated module over a Noetherian ring is also finitely generated.)
9. For k a field show that any subring of the polynomial ring $k[x]$ containing k is Noetherian. Give an example to show such subrings need not be U.F.D.s. [If $k \subset R \subseteq k[x]$ and $y \in R - k$ show that $k[x]$ is a finitely generated $k[y]$-module; then use the previous two exercises. For the second, consider $k[x^2, x^3]$.]
10. Prove that the subring $k[x, x^2y, x^3y^2, \dots, x^iy^{i-1}, \dots]$ of the polynomial ring $k[x, y]$ is *not* a Noetherian ring, hence not a finitely generated k-algebra. (Thus subrings of Noetherian rings need not be Noetherian and subalgebras of finitely generated k-algebras need not be finitely generated.)
11. Suppose R is a commutative ring in which all the prime ideals are finitely generated. This exercise proves that R is Noetherian.

(a) Prove that if the collection of ideals of R that are not finitely generated is nonempty, then it contains a maximal element I, and that R/I is a Noetherian ring.

(b) Prove that there are finitely generated ideals J_1 and J_2 containing I with $J_1J_2 \subseteq I$ and that J_1J_2 is finitely generated. [Observe that I is not a prime ideal.]

(c) Prove that I/J_1J_2 is a finitely generated R/I-submodule of J_1/J_1J_2. [Use Exercise 8.]

(d) Show that (c) implies the contradiction that I would be finitely generated over R and deduce that R is Noetherian.

12. Suppose R is a Noetherian ring and S is a finitely generated R-algebra. If $T \subseteq S$ is an R-algebra such that S is a finitely generated T*-module*, prove that T is a finitely generated R-algebra. [If $s_1, \dots, s_n$ generate S as an R-algebra, and $s'_1, \dots, s'_m$ generate S as a T-module, show that the elements s_i and $s'_j s'_k$ can be written as finite T-linear combinations of the s'_i. If T_0 is the R-subalgebra generated by the coefficients of these linear combinations, show S (hence T_0) is finitely generated (by the s'_i) as a T_0-module, and conclude that T is finitely generated as an R-algebra.]

13. Verify properties (1) to (10) of the maps $\mathcal{Z}$ and $\mathcal{I}$.

14. Show that the affine algebraic sets in $\mathbb{A}^1$ over any field k are $\emptyset$, k, and finite subsets of k.

15. If $k = \mathbb{F}_2$ and $V = \{(0,0), (1,1)\} \subset \mathbb{A}^2$, show that $\mathcal{I}(V)$ is the product ideal $\mathfrak{m}_1\mathfrak{m}_2$ where $\mathfrak{m}_1 = (x, y)$ and $\mathfrak{m}_2 = (x-1, y-1)$.

16. Suppose that V is a finite algebraic set in $\mathbb{A}^n$. If V has m points, prove that $k[V]$ is isomorphic as a k-algebra to k^m. [Use the Chinese Remainder Theorem.]

17. If k is a finite field show that every subset of $\mathbb{A}^n$ is an affine algebraic set.

18. If $k = \mathbb{F}_q$ is the finite field with q elements show that $\mathcal{I}(\mathbb{A}^1) = (x^q - x) \subset k[x]$.

19. For each nonconstant $f \in k[x]$ describe $\mathcal{Z}(f) \subseteq \mathbb{A}^1$ in terms of the unique factorization of f in $k[x]$, and then use this to describe $\mathcal{I}(\mathcal{Z}(f))$. Deduce that $\mathcal{I}(\mathcal{Z}(f)) = (f)$ if and only if f is the product of distinct linear factors in $k[x]$.

20. If f and g are irreducible polynomials in $k[x, y]$ that are not associates (do not divide each other), show that $\mathcal{Z}((f, g))$ is either $\emptyset$ or a finite set in $\mathbb{A}^2$. [If $(f, g) \neq (1)$, show (f, g) contains a nonzero polynomial in $k[x]$ (and similarly a nonzero polynomial in $k[y]$) by letting $R = k[x]$, $F = k(x)$, and applying Gauss's Lemma to show f and g are relatively prime in $F[y]$.]

21. Identify each 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with entries from k with the point (a, b, c, d) in $\mathbb{A}^4$. Show that the group $SL_2(k)$ of matrices of determinant 1 is an algebraic set in $\mathbb{A}^4$.

22. Prove that $SL_n(k)$ is an affine algebraic set in $\mathbb{A}^{n^2}$. [Generalize the preceding exercise.]

23. Let V be any line in $\mathbb{R}^2$ (the zero set of any nonzero linear polynomial $ax + by - c$). Prove that $\mathbb{R}[V]$ is isomorphic as an $\mathbb{R}$-algebra to the polynomial ring $\mathbb{R}[x]$, and give the corresponding isomorphism from $\mathbb{A}^1$ to V.

24. Let $V = \mathcal{Z}(xy - z) \subseteq \mathbb{A}^3$. Prove that V is isomorphic to $\mathbb{A}^2$ and provide an explicit isomorphism φ and associated k-algebra isomorphism $\widetilde{\varphi}$ from $k[V]$ to $k[\mathbb{A}^2]$, along with their inverses. Is $V = \mathcal{Z}(xy - z^2)$ isomorphic to $\mathbb{A}^2$?

25. Suppose $V \subseteq \mathbb{A}^n$ is an affine algebraic set and $f \in k[V]$. The *graph* of f is the collection of points $\{(a_1, \dots, a_n, f(a_1, \dots, a_n))\}$ in $\mathbb{A}^{n+1}$. Prove that the graph of f is an affine algebraic set isomorphic to V. [The morphism in one direction maps $(a_1, \dots, a_n)$ to $(a_1, \dots, a_n, f(a_1, \dots, a_n))$.]

26. Let $V = \mathcal{Z}(xz - y^2, yz - x^3, z^2 - x^2y) \subseteq \mathbb{A}^3$.

(a) Prove that the map $\varphi : \mathbb{A}^1 \to V$ defined by $\varphi(t) = (t^3, t^4, t^5)$ is a surjective morphism. [For the surjectivity, if $(x, y, z) \neq (0, 0, 0)$, let $t = y/x$.]

(b) Describe the corresponding k-algebra homomorphism $\widetilde{\varphi} : k[V] \to k[\mathbb{A}^1]$ explicitly.

(c) Prove that φ is not an isomorphism.

27. Suppose $\varphi : V \to W$ is a morphism of affine algebraic sets. If W' is an affine algebraic subset of W prove that the preimage $V' = \varphi^{-1}(W')$ of W' in V is an affine algebraic subset of V. If $W' = \mathcal{Z}(I)$ show that $V' = \mathcal{Z}(\widetilde{\varphi}(I))$ for the corresponding morphism $\widetilde{\varphi} : k[W] \to k[V]$.

28. Prove that if V and W are affine algebraic sets, then so is $V \times W$ and $k[V \times W] \cong k[V] \otimes_k k[W]$.

The following seven exercises introduce the notion of the *associated primes* of an R-module M. Cf. also Exercises 30–40 in Section 4 and Exercises 25–30 in Section 5.

Definition. A prime ideal P of R is said to be *associated* to the R-module M (sometimes called an *assassin* for M) if P is the annihilator of some element m of M, i.e., if M contains a submodule Rm isomorphic to R/P. The collection of associated primes for M is denoted $\operatorname{Ass}_R(M)$.

When $M = I$ is an ideal in R, it is customary to abuse the terminology and refer instead to the elements of $\operatorname{Ass}_R(R/I)$ (rather than the less interesting collection $\operatorname{Ass}_R(I)$) as the *primes associated to* I. (Cf. Exercises 28–29 in Section 5.)

29. If $R = \mathbb{Z}$ and $M = \mathbb{Z}/n\mathbb{Z}$, show that $\operatorname{Ass}_R(M)$ consists of the prime ideals (p) for the prime divisors p of n.

30. If M is the union of some collection of submodules M_i, prove that $\operatorname{Ass}_R(M)$ is the union of the collection $\operatorname{Ass}_R(M_i)$.

31. Suppose that $\operatorname{Ann}(m) = P$, i.e., that $Rm \cong R/P$. Prove that if $0 \neq m' \in Rm$ then $\operatorname{Ann}(m') = P$. Deduce that $\operatorname{Ass}_R(R/P) = \{P\}$. [Observe that R/P is an integral domain.]

32. Suppose that M is an R-module and that P is a maximal element in the collection of ideals of the form $\operatorname{Ann}(m)$, for $m \in M$. Prove that P is a prime ideal. [If $P = \operatorname{Ann}(m)$ and $ab \in P$, show that $bm \neq 0$ implies $\operatorname{Ann}(m) \subseteq \operatorname{Ann}(bm)$ and use the maximality of P to deduce that $a \in \operatorname{Ann}(bm) = P$.]

33. Suppose R is a Noetherian ring and $M \neq 0$ is an R-module. Prove that $\operatorname{Ass}_R(M) \neq \emptyset$. [Use Exercise 32.]

34. If L is a submodule of M with quotient $N \cong M/L$, prove that there are containments $\operatorname{Ass}_R(N) \subseteq \operatorname{Ass}_R(M) \subseteq \operatorname{Ass}_R(L) \cup \operatorname{Ass}_R(N)$, and show that both containments can be proper. [If $Rm \cong R/P$, show that $Rm \cap L = 0$ implies $P \in \operatorname{Ass}_R(N)$ and if $Rm \cap L \neq 0$ then $P \in \operatorname{Ass}_R(L)$ (by Exercise 31). For the second statement, consider $n\mathbb{Z} \subset \mathbb{Z}$.]

35. Suppose M is an R-module and let $\mathcal{S}$ be a subset of the prime ideals in $\operatorname{Ass}_R(M)$. Prove there is a submodule N of M with $\operatorname{Ass}_R(N) = \mathcal{S}$ and $\operatorname{Ass}_R(M/N) = \operatorname{Ass}_R(M) - \mathcal{S}$. [Consider the collection of submodules N' of M with $\operatorname{Ass}_R(N') \subseteq \mathcal{S}$. Use Exercise 30 and Zorn's Lemma to show that there is a maximal submodule N subject to $\operatorname{Ass}_R(N) \subseteq \mathcal{S}$. If $P \in \operatorname{Ass}_R(M/N)$, there is a submodule $M'/N \cong R/P$. Use the previous exercise to show that $\operatorname{Ass}_R(M') \subseteq \operatorname{Ass}_R(N) \cup \{P\}$ and then use maximality of N to show $P \in \operatorname{Ass}_R(M) - \mathcal{S}$, so that $\operatorname{Ass}_R(M/N) \subseteq \operatorname{Ass}_R(M) - \mathcal{S}$ and $\operatorname{Ass}_R(N) \subseteq \mathcal{S}$. Use the previous exercise again to conclude that equality holds in each.]

Suppose M is a finitely generated module over the commutative ring R with generators $m_1, \dots, m_n$. The *Fitting ideal* $\mathcal{F}_R(M)$ (of level 0) of M (also called a *determinant ideal*) is the ideal in R generated by the determinants of all $n \times n$ matrices $A = (r_{ij})$ where $r_{ij} \in R$ and $r_{i1}m_1 + \cdots + r_{in}m_n = 0$ in M, i.e., the rows of A consist of the coefficients in R of relations among the generators m_i (A is called an $n \times n$ "relations matrix" for M). The following five exercises outline some of the properties of the Fitting ideal.

36. **(a)** Show that the Fitting ideal of M is also the ideal in R generated by all the $n \times n$ minors of all $p \times n$ matrices $A = (r_{ij})$ for $p \geq 1$ whose rows consist of the coefficients in R of relations among the generators m_i.

(b) Let A be a fixed $p \times n$ matrix as in (a) and let A' be a $p \times n$ matrix obtained from A by any elementary row or column operation. Show that the ideal in R generated by all the $n \times n$ minors of A is the same as the ideal in R generated by all the $n \times n$ minors of A'.

37. Suppose $m_1, \dots, m_n$ and $m'_1, \dots, m'_{n'}$ are two sets of R-module generators for M. Let $\mathcal{F}$ denote the Fitting ideal for M computed using the generators $m_1, \dots, m_n$ and let $\mathcal{F}'$ denote the Fitting ideal for M computed using the generators $m_1, \dots, m_n, m'_1, \dots, m'_{n'}$.

(a) Show that $m'_s = a_{s'1}m_1 + \cdots + a_{s'n}m_n$ for some $a_{s'1}, \dots, a_{s'n} \in R$, and deduce that $(-a_{s'1}, \dots, -a_{s'n}, 0, \dots, 0, 1, 0, \dots 0)$ is a relation among $m_1, \dots, m_n, m'_1, \dots, m'_{n'}$.

(b) If $A = (r_{ij})$ is an $n \times n$ matrix whose rows are the coefficients of relations among $m_1, \dots, m_n$ show that $\det A = \det A'$ where A' is an $(n+n') \times (n+n')$ matrix whose rows are the coefficients of relations among $m_1, \dots, m_n, m'_1, \dots, m'_{n'}$. Deduce that $\mathcal{F} \subseteq \mathcal{F}'$. [Use (a) to find a block upper triangular A' having A in the upper left block and the $n' \times n'$ identity matrix in the lower right block.]

(c) Prove that $\mathcal{F}' \subseteq \mathcal{F}$ and conclude that $\mathcal{F}' = \mathcal{F}$. [Use the previous exercise.]

(d) Deduce from (c) that the Fitting ideal $\mathcal{F}_R(M)$ of M is an invariant of M that does not depend on the choice of generators for M used to compute it.

38. All modules in this exercise are assumed finitely generated.

(a) If M can be generated by n elements prove that $\text{Ann}(M)^n \subseteq \mathcal{F}_R(M) \subseteq \text{Ann}(M)$, where $\text{Ann}(M)$ is the annihilator of M in R. [If A is an $n \times n$ relations matrix for M, then $AX = 0$, where X is the column matrix whose entries are $m_1, \dots, m_n$. Multiply by the adjoint of A to deduce that $\det A$ annihilates M.]

(b) If $M = M_1 \times M_2$ is the direct product of the R-modules M_1 and M_2 prove that $\mathcal{F}_R(M) = \mathcal{F}_R(M_1)\mathcal{F}_R(M_2)$.

(c) If $M = (R/I_1) \times \cdots \times (R/I_n)$ is the direct product of cyclic R-modules for ideals I_i in R prove that $\mathcal{F}_R(M) = I_1 I_2 \dots I_n$.

(d) If $R = \mathbb{Z}$ and M is a finitely generated abelian group show that $\mathcal{F}_{\mathbb{Z}}(M) = 0$ if M is infinite and $\mathcal{F}_{\mathbb{Z}}(M) = |M|\mathbb{Z}$ if M is finite.

(e) If I is an ideal in R prove that the image of $\mathcal{F}_R(M)$ in the quotient R/I is $\mathcal{F}_{R/I}(M/IM)$.

(f) Prove that $\mathcal{F}_R(M/IM) \subseteq (\mathcal{F}_R(M), I) \subseteq R$.

(g) If $\varphi : M \to M'$ is a surjective R-module homomorphism prove $\mathcal{F}_R(M) \subseteq \mathcal{F}_R(M')$.

(h) If $0 \to L \to M \to N \to 0$ is a short exact sequence of R-modules, prove that $\mathcal{F}_R(L)\mathcal{F}_R(N) \subseteq \mathcal{F}_R(M)$.

(i) Suppose R is the polynomial ring $k[x, y, z]$ over the field k. Let $M = R/(x, y^2, yz, z^2)$ and let L be the submodule $(x, y, z)/(x, y^2, yz, z^2)$ of M. Prove that $\mathcal{F}_R(M)$ is (x, y^2, yz, z^2) and $\mathcal{F}_R(L)$ is $(x, y, z)^2$. (This shows that in general the Fitting ideal of a submodule L of M need not contain the Fitting ideal for M.)

39. Suppose M is an R-module and that $\varphi : R^n \to M$ is a surjective R-module homomorphism (i.e., M can be generated by n elements). Let $L = \ker \varphi$. Prove that the image of the

R-module homomorphism from $\bigwedge^n(L) \to \bigwedge^n(R^n) \cong R$ induced by the inclusion of L in R^n is the Fitting ideal $\mathcal{F}_R(M)$.

40. Suppose R and S are commutative rings, $\varphi : R \to S$ is a ring homomorphism, M is a finitely generated R-module, and $M' = S \otimes_R M$ is the S-module obtained by extending scalars from R to S. Prove that the Fitting ideal $\mathcal{F}_S(M')$ for M' over S is the extension to S of the Fitting ideal $\mathcal{F}_R(M)$ for M over R.

The following two exercises indicate how the remainder in Theorem 23 of Chapter 9 can be used to effect computations in quotients of polynomial rings.

41. Suppose $\{g_1, \dots, g_m\}$ is a Gröbner basis for the ideal I in $k[x_1, \dots, x_n]$. Prove that the monomials m not divisible by any $LT(g_i)$, $1 \le i \le m$, give a k-vector space basis for the quotient $k[x_1, \dots, x_n]/I$.

42. Let $I = (x^3y - xy^2 + 1, x^2y^2 - y^3 - 1)$ as in Example 1 following Proposition 9.26.

(a) Use the previous exercise to show that $\{1, y, y^2, y^3\}$ is a basis for the k-vector space $k[x, y]/I$.

(b) Compute the 4×4 multiplication table for the basis vectors in (a).

43. Suppose $K[x_1, \dots, x_n]$ is a polynomial ring in n variables over a field K and k is a subfield of K. If I is an ideal in $k[x_1, \dots, x_n]$, let I' be the ideal generated by I in $K[x_1, \dots, x_n]$.

(a) If G is a Gröbner basis for the ideal I in $k[x_1, \dots, x_n]$ with respect to some monomial ordering, show that G is also a Gröbner basis for the ideal I' in $K[x_1, \dots, x_n]$ with respect to the same monomial ordering. [Use Buchberger's Criterion.]

(b) Prove that the dimension of the quotient $k[x_1, \dots, x_n]/I$ as a vector space over k is the same as the dimension of the quotient $K[x_1, \dots, x_n]/I'$ as a vector space over K. [One method: use (a) and Exercise 41.]

(c) Prove that $I = k[x_1, \dots, x_n]$ if and only if $I' = K[x_1, \dots, x_n]$.

44. Let $V = \mathcal{Z}(x^3 - x^2z - y^2z)$ and $W = \mathcal{Z}(x^2 + y^2 - z^2)$ in $\mathbb{C}^3$. Then $\mathcal{I}(V) = (x^3 - x^2z - y^2z)$ and $\mathcal{I}(W) = (x^2 + y^2 - z^2)$ in $\mathbb{C}[x, y, z]$ (cf. Exercise 23 in Section 3). Show that $\varphi((a, b, c)) = (a^2c - b^2c, 2abc, -a^3)$ defines a morphism from V to W.

45. Let $V = \mathcal{Z}(x^3 + y^3 + 7z^3) \subset \mathbb{C}^3$. Then $\mathcal{I}(V) = (x^3 + y^3 + 7z^3)$ in $\mathbb{C}[x, y, z]$ (cf. Exercise 24 in Section 3).

(a) Show that

$$\widetilde{\varphi}(x) = x(y^3 - 7z^3), \qquad \widetilde{\varphi}(y) = y(7z^3 - x^3), \qquad \widetilde{\varphi}(z) = z(x^3 - y^3)$$

defines a $\mathbb{C}$-algebra homomorphism from $k[V]$ to itself.

(b) Let $\varphi : V \to V$ be the morphism corresponding to $\widetilde{\varphi}$. Observe that $(-2, 1, 1) \in V$ and compute $\varphi((-2, 1, 1)) \in V$.

(c) Prove there are infinitely many points (a, b, c) on V with $a, b, c \in \mathbb{Z}$ and the greatest common divisor of a, b, and c is 1.

46. Let $V = \mathcal{Z}(xz + y^2 + z^2, xy - xz + yz - 2z^2) \subset \mathbb{C}^3$ and $W = \mathcal{Z}(u^3 - uv^2 + v^3) \subset \mathbb{C}^2$ as in Example 2 following Corollary 9. Show that the map $\varphi((a, b)) = (-2a^2 + ab, ab - b^2, a^2 - ab)$ defines a morphism from W to V. Show the corresponding $\mathbb{C}$-algebra homomorphism from $k[V]$ to $k[W]$ has a kernel generated by $x^2 - 3y^2 + yz$.

47. Define $\Phi : \mathbb{Q}[u, v, w] \to \mathbb{Q}[x, y]$ by $\Phi(u) = x^2 + y$, $\Phi(v) = x + y^2$, and $\Phi(w) = x - y$. Show that neither x nor y is in the image of Φ. Show that $f = 2x^3 - 4xy - 2y^3 - 4y$ is in the image of Φ and find a polynomial in $\mathbb{Q}[u, v, w]$ mapping to f. Show that $\ker \Phi$ is the ideal generated by

$$u^2 - 2uv - 2uw^2 + 4uw + v^2 - 2vw^2 - 4vw + w^4 + 3w^2.$$

48. Suppose α is a root of the irreducible polynomial $p(x) \in k[x]$ and $\beta = f(\alpha)/g(\alpha)$ with polynomials $f(x), g(x) \in k[x]$ where $g(\alpha) \neq 0$.

- **(a)** Show $ag + bp = 1$ for some polynomials $a, b \in k[x]$ and show $\beta = h(\alpha)$ where $h = af$.
- **(b)** Show that the ideals $(p, y - h)$ and $(p, gy - f)$ are equal in $k[x, y]$.
- **(c)** Conclude that the minimal polynomial for β is the monic polynomial in $G \cap k[y]$ where G is the reduced Gröbner basis for the ideal $(p, gy - f)$ in $k[x, y]$ for the lexicographic monomial ordering $x > y$.
- **(d)** Find the minimal polynomial over $\mathbb{Q}$ of $(3 - \sqrt[3]{2} + \sqrt[3]{4})/(1 + 3\sqrt[3]{2} - 3\sqrt[3]{4})$.

15.2 RADICALS AND AFFINE VARIETIES

Since the zeros of a polynomial f are the same as the zeros of the powers $f^2, f^3, \ldots$ in general there are many different ideals in the ring $k[x_1, x_2, \ldots, x_n]$ whose zero locus define the same algebraic set V in affine n-space. This leads to the notion of the radical of an ideal, which can be defined in any commutative ring:

Definition. Let I be an ideal in a commutative ring R.

- **(1)** The *radical* of I, denoted by $\operatorname{rad} I$, is the collection of elements in R some power of which lie in I, i.e.,
$$\operatorname{rad} I = \{a \in R \mid a^k \in I \text{ for some } k \geq 1\}.$$
- **(2)** The radical of the zero ideal is called the *nilradical* of R.
- **(3)** An ideal I is called a *radical* ideal if $I = \operatorname{rad} I$.

Note that $a \in R$ is in the nilradical of R if and only if some power of a is 0, so the nilradical of R is the set of all nilpotent elements of R.

Proposition 11. Let I be an ideal in the commutative ring R. Then $\operatorname{rad} I$ is an ideal containing I, and $(\operatorname{rad} I)/I$ is the nilradical of R/I. In particular, R/I has no nilpotent elements if and only if $I = \operatorname{rad} I$ is a radical ideal.

Proof: It is clear that $I \subseteq \operatorname{rad} I$. By definition, the nilradical of R/I consists of the elements in the quotient some power of which is 0. Under the Lattice Isomorphism Theorem for rings this collection of elements corresponds to the elements of R some power of which lie in I, i.e., $\operatorname{rad} I$. It is therefore sufficient to prove that the nilradical N of any commutative ring R is an ideal. Since $0 \in N$, $N \neq \emptyset$. If $a \in N$ and $r \in R$, then since $a^n = 0$ for some $n \geq 1$, the commutativity of R implies that $(ra)^n = r^n a^n = 0$, so $ra \in N$. It remains to see that if $a, b \in N$ then $a + b \in N$. Suppose $a^n = 0$ and $b^m = 0$. Since the Binomial Theorem holds in the commutative ring R (cf. Exercise 25 in Section 7.3),

$$(a + b)^{n+m} = \sum_{i=0}^{n+m} r_i a^i b^{n+m-i}$$

for some ring elements r_i (the binomial coefficients in R). For each term in this sum either $i \geq n$ (in which case $a^i = 0$) or $n + m - i \geq m$, (in which case $b^{n+m-i} = 0$). Hence $(a + b)^{n+m} = 0$, which shows that $a + b$ is nilpotent, i.e., $a + b \in N$.

Proposition 12. The radical of a proper ideal I is the intersection of all prime ideals containing I. In particular, the nilradical is the intersection of all the prime ideals in R.

Proof: Passing to R/I, Proposition 11 shows that it suffices to prove this result for $I = 0$, and in this case the statement is that the nilradical N of R is the intersection of all the prime ideals in R. Let N' denote the intersection of all the prime ideals in R.

Let a be any nilpotent element in R and let P be any prime ideal. Since $a^k = 0$ for some k, there is a smallest positive power n such that $a^n \in P$. Then the product $a^{n-1}a \in P$, and since P is prime, either $a^{n-1} \in P$ or $a \in P$. The former contradicts the minimality of n, and so $a \in P$. Since P was arbitrary, $a \in N'$, which shows that $N \subseteq N'$.

We prove the reverse containment $N' \subseteq N$ by showing that if $a \notin N$, then $a \notin N'$. If a is an element of R not contained in N, let $\mathcal{S}$ be the family of all proper ideals not containing any positive power of a. The collection $\mathcal{S}$ is not empty since $0 \in \mathcal{S}$. Also, if a^k is not contained in any ideal in the chain $I_1 \subseteq I_2 \subseteq \cdots$, then a^k is also not contained in the union of these ideals, which shows that chains in $\mathcal{S}$ have upper bounds. By Zorn's Lemma, $\mathcal{S}$ has a maximal element, P. The ideal P must in fact be a prime ideal, as follows. Suppose for some x and y not contained in P, the product xy is an element of P. By the maximality of P, $a^n \in (x) + P$ and $a^m \in (y) + P$ for some positive integers n and m. Then $a^{n+m} \in (xy) + P = P$ contradicting the fact that P is an element of $\mathcal{S}$. This shows that P is indeed a prime ideal not containing a, and hence $a \notin N'$, completing the proof.

Note that in Noetherian rings, Theorem 2 can be used to circumvent the appeal to Zorn's Lemma in the preceding proof.

Corollary 13. Prime (and hence also maximal) ideals are radical.

Proof: If P is a prime ideal, then P is clearly the intersection of all the prime ideals containing P, so $P = \text{rad } P$ by the proposition.

Examples

(1) In the ring of integers $\mathbb{Z}$, the ideal (a) is a radical ideal if and only if a is squarefree or zero. More generally, if $a = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$ with $a_i \geq 1$ for all i, is the prime factorization of the positive integer a, then $\text{rad}(a) = (p_1 p_2 \cdots p_r)$. For instance, $\text{rad}(180) = (30)$. Note that $(p_1), (p_2), \ldots, (p_r)$ are precisely the prime ideals containing the ideal (a) and that their intersection is the ideal $(p_1 p_2 \cdots p_r)$. More generally, in any U.F.D. R, $\text{rad}(a) = (p_1 p_2 \cdots p_r)$ if $a = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$ is the unique factorization of a into distinct irreducibles.

(2) The ideal $(x^3 - y^2)$ in $k[x, y]$ is a prime ideal (Exercise 14, Section 9.1), hence is radical.

(3) If $l_1, \ldots, l_m$ are linear polynomials in $k[x_1, x_2, \ldots, x_n]$ then $I = (l_1, \ldots, l_m)$ is either $k[x_1, x_2, \ldots, x_n]$ or a prime ideal, hence I is a radical ideal.

Proposition 14. If R is a Noetherian ring then for any ideal I some positive power of $\text{rad } I$ is contained in I. In particular, the nilradical, N, of a Noetherian ring is a nilpotent ideal: $N^k = 0$ for some $k \geq 1$.

Proof: For any ideal I, the ideal rad I is finitely generated since R is Noetherian. If $a_1, \dots, a_m$ are generators of rad I, then by definition of the radical, for each i we have $a_i^{k_i} \in I$ for some positive integer k_i. Let k be the maximum of all the k_i. Then the ideal $(\text{rad}\, I)^{km}$ is generated by elements of the form $a_1^{d_1} a_2^{d_2} \cdots a_m^{d_m}$ where $d_1 + \cdots + d_m = km$, and each of these elements has at least one factor $a_i^{d_i}$ with $d_i \geq k$. Then $a_i^{d_i} \in I$, hence each generator of $(\text{rad}\, I)^{km}$ lies in I, and so $(\text{rad}\, I)^{km} \subseteq I$.

The Zariski Topology

We saw in the preceding section that if we restrict to the set of ideals I of $k[\mathbb{A}^n]$ arising as the ideals associated with some algebraic set V, i.e., with $I = \mathcal{I}(V)$, then the maps $\mathcal{Z}$ (from such ideals to algebraic sets) and $\mathcal{I}$ (from algebraic sets to ideals) are inverses of each other: $\mathcal{Z}(\mathcal{I}(V)) = V$ and $\mathcal{I}(\mathcal{Z}(I)) = I$. The elements of the ring $k[\mathbb{A}^n]/\mathcal{I}(V)$ give k-valued functions on V and, since k has no nilpotent elements, powers of nonzero functions are also nonzero functions. Put another way, the ring $k[\mathbb{A}^n]/\mathcal{I}(V)$ has no nilpotent elements, so by Proposition 11, the ideal $\mathcal{I}(V)$ is always a radical ideal.

For arbitrary fields k, it is in general not true that every radical ideal is the ideal of some algebraic set, i.e., of the form $\mathcal{I}(V)$ for some algebraic set V. For example, the ideal $(x^2 + 1)$ in $\mathbb{R}[x]$ is maximal, hence is a radical ideal (by Corollary 13), but is not the ideal of any algebraic set — if it were, then $x^2 + 1$ would have to vanish on that set, but $x^2 + 1$ has no zeros in $\mathbb{R}$. A similar construction works for any field k that is not algebraically closed — there exists an irreducible polynomial $p(x)$ of degree at least 2 in $k[x]$, which then generates the maximal (hence radical) ideal $(p(x))$ in $k[x]$ that has no zeros in k. It is perhaps surprising that the presence of polynomials in one variable that have no zeros is the *only* obstruction to a radical ideal (in *any* number of variables) not being of the form $\mathcal{I}(V)$. This is shown by the next theorem, which provides a fundamental connection between "geometry" and "algebra" and shows that over an *algebraically closed* field (such as $\mathbb{C}$) every radical ideal is of the form $\mathcal{I}(V)$. Over these fields the "geometrically defined" ideals $I = \mathcal{I}(V)$ are therefore the same as the radical ideals, which is a "purely algebraic" property of the ideal I (namely that $I = \text{rad}\, I$).

Theorem. *(Hilbert's Nullstellensatz)* Let E be an algebraically closed field. Then $\mathcal{I}(\mathcal{Z}(I)) = \text{rad}\, I$ for every ideal I of $E[x_1, x_2, \dots, x_n]$. Moreover, the maps $\mathcal{Z}$ and $\mathcal{I}$ in the correspondence

$$\begin{tikzcd}
\{\text{affine algebraic sets}\} \arrow[r, shift left, "\mathcal{I}"] & \{\text{radical ideals}\} \arrow[l, shift left, "\mathcal{Z}"]
\end{tikzcd}$$

are bijections that are inverses of each other.

Proof: This will be proved in the next section (cf. Theorem 32).

Example

The maps $\mathcal{I}$ and $\mathcal{Z}$ in the Nullstellensatz are defined over any field k, and as mentioned are not bijections if k is not algebraically closed. For any field k, however, the map $\mathcal{Z}$ is always surjective and the map $\mathcal{I}$ is always injective (cf. Exercise 9).

One particular consequence of the Nullstellensatz is that for any *proper* ideal I we have $\mathcal{Z}(I) \neq \emptyset$ since $\text{rad}\, I \neq k[\mathbb{A}^n]$. Hence there always exists at least one common zero ("nullstellen" in German) for all the polynomials contained in a proper ideal (over an algebraically closed field).

We next see that the affine algebraic sets define a topology on affine n-space. Recall that a *topological space* is any set X together with a collection of subsets $\mathcal{T}$ of X, called the *closed sets* in X, satisfying the following axioms:

(i) an arbitrary intersection of closed sets is closed: if $S_i \in \mathcal{T}$ for i in any index set, then $\cap S_i \in \mathcal{T}$,

(ii) a finite union of closed sets is closed: if $S_1, \dots, S_q \in \mathcal{T}$ then $S_1 \cup \cdots \cup S_q \in \mathcal{T}$, and

(iii) the empty set and the whole space are closed: $\emptyset, X \in \mathcal{T}$.

A subset U of X is called *open* if its complement, $X - U$, is closed (i.e., $X - U \in \mathcal{T}$). The axioms for a topological space are often (equivalently) phrased in terms of the collection of open sets in X.

There are many examples of topological spaces, and a wealth of books on topology. A fixed set X may have a number of different topologies on it, and the collections of closed sets need not be related in these different structures. On any set X there are always at least two topologies: the so-called discrete topology in which every subset of X is closed (i.e., $\mathcal{T}$ is the collection of *all* subsets of X), and the so-called trivial topology in which the only closed sets are $\emptyset$ and X required by axiom (iii).

Suppose now that $X = \mathbb{A}^n$ is affine n-space over an arbitrary field k. Then the collection $\mathcal{T}$ consisting of all the affine algebraic sets in $\mathbb{A}^n$ satisfies the three axioms for a topological space — these are precisely properties (3), (4) and (5) of algebraic sets in the preceding section. It follows that these sets can be taken to be the closed sets in a topology on $\mathbb{A}^n$:

Definition. The *Zariski topology* on affine n-space over an arbitrary field k is the topology in which the closed sets are the affine algebraic sets in $\mathbb{A}^n$.

The Zariski topology is quite "coarse" in the sense that there are "relatively few" closed (or open) sets. For example, for the Zariski topology on $\mathbb{A}^1$ the only closed sets are $\emptyset$, k and the finite sets (cf. Exercise 14 in Section 1), and so the nonempty open sets are the complements of finite sets. If k is an infinite field it follows that in the Zariski topology any two nonempty open sets in $\mathbb{A}^1$ have nonempty intersection. In the language of point-set topology, the Zariski topology is always T_1 (points are closed sets), but for infinite fields the Zariski topology is never T_2 (Hausdorff), i.e., two distinct points never belong to two disjoint open sets (cf. the exercises). For example, when $k = \mathbb{R}$, a nonempty Zariski open set is just the real line $\mathbb{R}$ with some finite number of points removed, and any two such sets have (infinitely many) points in common. Note also that the Zariski open (respectively, closed) sets in $\mathbb{R}$ are also open (respectively, closed) sets with respect to the usual Euclidean topology. The converse is not true; for example the interval $[0,1]$ is closed in the Euclidean topology but is not closed in the Zariski topology. In this sense the Euclidean topology on $\mathbb{R}$ is much "finer"; there are many more open sets in the Euclidean topology, in fact the collection of Euclidean open (respectively, closed) sets properly contains the collection of Zariski open (respectively, closed) sets.

The Zariski topology on $\mathbb{A}^n$ is defined so that the affine algebraic subsets of $\mathbb{A}^n$ are closed. In other words, the topology is defined by the zero sets of the ideals in the coordinate ring of $\mathbb{A}^n$. A similar definition can be used to define a Zariski topology on *any* algebraic set V in $\mathbb{A}^n$, as follows. If $k[V]$ is the coordinate ring of V, then the distinct elements of $k[V]$ define distinct k-valued functions on V and there is a natural way of defining

$$\begin{aligned} \mathcal{Z} &: \{\text{ ideals in } k[V]\,\} \longrightarrow \{\text{ algebraic subsets of } V\,\} \\ \mathcal{I} &: \{\text{ subsets of } V\,\} \quad \longrightarrow \{\text{ ideals in } k[V]\,\} \end{aligned}$$

just as for the case $V = \mathbb{A}^n$. For example, if $\overline{J}$ is an ideal in $k[V]$, then $\mathcal{Z}(\overline{J})$ is the set of elements in V that are common zeros of all the functions in the ideal $\overline{J}$. It is easy to verify that the resulting zero sets in V satisfy the three axioms for a topological space, defining a *Zariski topology on* V, where the closed sets are the algebraic subsets, $\mathcal{Z}(\overline{J})$, for any ideal $\overline{J}$ of $k[V]$. By the Lattice Isomorphism Theorem, the ideals of $k[V]$ are the ideals of $k[x_1, \dots, x_n]$ that contain $\mathcal{I}(V)$ taken mod $\mathcal{I}(V)$. If J is the complete preimage in $k[x_1, \dots, x_n]$ of $\overline{J}$, then the locus of J in $\mathbb{A}^n$ is the same as the locus of $\overline{J}$ in V. It follows that this definition of the Zariski topology on V is just the *subspace topology* for $V \subseteq \mathbb{A}^n$. (Recall that in a topological space X, the closed sets with respect to the subspace topology of a subspace Y are defined to be the sets $C \cap Y$, where C is a closed set in X.) The advantage to the definition of the Zariski topology on V above is that it is defined intrinsically in terms of the coordinate ring $k[V]$ of V, and since the isomorphism type of $k[V]$ does not depend on the affine space $\mathbb{A}^n$ containing V, the Zariski topology on V also depends only on V and not on the ambient affine space in which V may be embedded.

If V and W are two affine algebraic spaces, then since a morphism $\varphi : V \to W$ is defined by polynomial functions, it is easy to see that φ is *continuous* with respect to the Zariski topologies on V and W (cf. Exercise 27 in Section 1, which shows that the inverse image of a Zariski closed set under a morphism is Zariski closed). In fact the Zariski topology is the coarsest topology in which points are closed and for which polynomial maps are continuous. There exist maps that are continuous with respect to the Zariski topology that are not morphisms, however (cf. Exercise 17).

We have the usual topological notions of closure and density with respect to the Zariski topology.

Definition. For any subset A of $\mathbb{A}^n$, the *Zariski closure* of A is the smallest algebraic set containing A. If $A \subseteq V$ for an algebraic set V then A is *Zariski dense* in V if the Zariski closure of A is V.

For example, if $k = \mathbb{R}$, the algebraic sets in $\mathbb{A}^1$ are $\emptyset$, $\mathbb{R}$, and finite subsets of $\mathbb{R}$ by Exercise 14 in Section 1. The Zariski closure of any infinite set A of real numbers is then all of $\mathbb{A}^1$ and A is Zariski dense in $\mathbb{A}^1$.

Proposition 15. The Zariski closure of a subset A in $\mathbb{A}^n$ is $\mathcal{Z}(\mathcal{I}(A))$.

Proof: Certainly $A \subseteq \mathcal{Z}(\mathcal{I}(A))$. Suppose V is any algebraic set containing A: $A \subseteq V$. Then $\mathcal{I}(V) \subseteq \mathcal{I}(A)$ and $\mathcal{Z}(\mathcal{I}(A)) \subseteq \mathcal{Z}(\mathcal{I}(V)) = V$, so $\mathcal{Z}(\mathcal{I}(A))$ is the smallest algebraic set containing A.

If $\varphi : V \to W$ is a morphism of algebraic sets, the image $\varphi(V)$ of V need not be an algebraic subset of W, i.e., need not be Zariski closed in W. For example the projection of the hyperbola $V = \mathcal{Z}(xy - 1)$ in $\mathbb{R}^2$ onto the x-axis has image $\mathbb{R}^1 - \{0\}$, which as we have just seen is not an affine algebraic set.

The next result shows that the Zariski closure of the image of a morphism is determined by the kernel of the associated k-algebra homomorphism.

Proposition 16. Suppose $\varphi : V \to W$ is a morphism of algebraic sets and $\widetilde{\varphi} : k[W] \to k[V]$ is the associated k-algebra homomorphism of coordinate rings. Then

(1) The kernel of $\widetilde{\varphi}$ is $\mathcal{I}(\varphi(V))$.

(2) The Zariski closure of $\varphi(V)$ is the zero set in W of $\ker \widetilde{\varphi}$. In particular, the homomorphism $\widetilde{\varphi}$ is injective if and only if $\varphi(V)$ is Zariski dense in W.

Proof: Since $\widetilde{\varphi} = f \circ \varphi$, we have $\widetilde{\varphi}(f) = 0$ if and only if $(f \circ \varphi)(P) = 0$ for all $P \in V$, i.e., $f(Q) = 0$ for all $Q = \varphi(P) \in \varphi(V)$, which is the statement that $f \in \mathcal{I}(\varphi(V))$, proving the first statement. Since the Zariski closure of $\varphi(V)$ is the zero set of $\mathcal{I}(\varphi(V))$ by the previous proposition, the first statement in (2) follows.

If $\widetilde{\varphi}$ is injective then the Zariski closure of $\varphi(V)$ is $\mathcal{Z}(0) = W$ and so $\varphi(V)$ is Zariski dense. Conversely, suppose $\varphi(V)$ is Zariski dense in W, i.e., $\mathcal{Z}(\mathcal{I}(\varphi(V))) = W$. Then $\mathcal{I}(\varphi(V)) = \mathcal{I}(\mathcal{Z}(\mathcal{I}(\varphi(V)))) = \mathcal{I}(W) = 0$ and so $\ker \widetilde{\varphi} = 0$.

By Proposition 16 the ideal of polynomials defining the Zariski closure of the image of a morphism φ is the kernel of the corresponding k-algebra homomorphism $\widetilde{\varphi}$ in Theorem 6. Proposition 8(1) allows us to compute this kernel using Gröbner bases.

Example: (Implicitization)

A morphism $\varphi : \mathbb{A}^n \to \mathbb{A}^m$ is just a map

$$\varphi((a_1, a_2, \dots, a_n)) = (\varphi_1(a_1, a_2, \dots, a_n), \dots, \varphi_m(a_1, a_2, \dots, a_n))$$

where φ_i is a polynomial. If k is an infinite field, then $\mathcal{I}(\mathbb{A}^m)$ and $\mathcal{I}(\mathbb{A}^n)$ are both 0, so we may write $k[\mathbb{A}^m] = k[y_1, \dots, y_m]$ and $k[\mathbb{A}^n] = k[x_1, \dots, x_n]$. The k-algebra homomorphism $\widetilde{\varphi} : k[\mathbb{A}^m] \to k[\mathbb{A}^n]$ corresponding to φ is then defined by mapping y_i to $\varphi_i = \varphi_i(x_1, \dots, x_n)$. The image $\varphi(\mathbb{A}^n)$ consists of the set of points $(b_1, \dots, b_m)$ with

$$\begin{aligned} b_1 &= \varphi_1(a_1, a_2, \dots, a_n) \\ b_2 &= \varphi_2(a_1, a_2, \dots, a_n) \\ &\vdots \\ b_m &= \varphi_m(a_1, a_2, \dots, a_n) \end{aligned}$$

where $a_i \in k$. This is the collection of points in $\mathbb{A}^m$ *parametrized* by the functions $\varphi_1, \dots, \varphi_m$ (with the a_i as parameters). In general such a parametrized collection of points is not an algebraic set. Finding the equations for the smallest algebraic set containing these points is referred to as *implicitization,* since it amounts to finding a ('smallest') collection of equations satisfied by the b_i (the 'implicit' algebraic relations).

By Proposition 16, this algebraic set is the Zariski closure of $\varphi(\mathbb{A}^n)$ and is the zero set of $\ker\widetilde{\varphi}$. By Proposition 8 this kernel is given by $\mathcal{A} \cap k[y_1, \dots, y_m]$, where $\mathcal{A}$ is the ideal in $k[x_1, \dots, x_n, y_1, \dots, y_m]$ generated by the polynomials $y_1 - \varphi_1, \dots, y_m - \varphi_m$. If we compute the reduced Gröbner basis G for $\mathcal{A}$ with respect to the lexicographic monomial ordering $x_1 > \cdots > x_n > y_1 > \cdots > y_m$, then the polynomials of G lying in $k[y_1, \dots, y_m]$ generate $\ker\tilde{\pi}$. The zero set of these polynomials defines the Zariski closure of $\varphi(\mathbb{A}^n)$ and therefore give the implicitization.

For an explicit example, consider the points $A = \{(a^2, a^3) \mid a \in \mathbb{R}\}$ in $\mathbb{R}^2$. Using coordinates x, y for $\mathbb{R}^2$ and t for $\mathbb{R}^1$, the ideal $\mathcal{A}$ in $\mathbb{R}[x, y, z, t]$ is $(x - t^2, y - t^3)$. The only element of the reduced Gröbner basis for $\mathcal{A}$ for the ordering $t > x > y$ lying in $\mathbb{R}[x, y]$ is $x^3 - y^2$, so $\mathcal{Z}(x^3 - y^2)$ is the smallest algebraic set in $\mathbb{R}^2$ containing A.

Example: (Projections of Algebraic Sets)

Suppose $V \subseteq \mathbb{A}^n$ is an algebraic set and $m < n$. Let $\pi : V \to \mathbb{A}^m$ be the morphism projecting onto the first m coordinates:

$$\pi((a_1, a_2, \dots, a_n)) = (a_1, a_2, \dots, a_m).$$

If we use coordinates $x_1, \dots, x_n$ in $k[V]$ and coordinates $y_1, \dots, y_m$ in $k[\mathbb{A}^m]$, the k-algebra homomorphism corresponding to π is given by the map

$$\begin{aligned}\tilde{\pi} : k[y_1, \dots, y_m] &\longrightarrow k[x_1, \dots, x_n]/\mathcal{I}(V)\\ y_i &\longmapsto x_i.\end{aligned}$$

Suppose $V = \mathcal{Z}(I)$ and $I = (f_1, \dots, f_s)$. The Zariski closure of $\pi(V)$ is the zero set of $\ker\tilde{\pi} = \mathcal{A} \cap k[y_1, \dots, y_m]$ where $\mathcal{A}$ is the ideal in $k[x_1, \dots, x_n, y_1, \dots, y_m]$ generated by the polynomials $y_1 - x_1, \dots, y_m - x_m$ together with a set of generators for $\mathcal{I}(V)$. The polynomials involving only $y_1, \dots, y_m$ in the reduced Gröbner basis G for $\mathcal{A}$ with respect to the lexicographic monomial ordering $x_1 > \cdots > x_n > y_1 > \cdots > y_m$ are generators for the Zariski closure of $\pi(V)$.

If k is algebraically closed we can actually do better with the help of the Nullstellensatz, which gives $\mathcal{I}(V) = \operatorname{rad} I$. Then it is straightforward to see that we obtain the same zero set if in the ideal $\mathcal{A}$ we replace the generators for $\mathcal{I}(V)$ by the generators $f_1, \dots, f_s$ of I (cf. Exercise 46).

For an explicit example, consider projection onto the first two coordinates of $V = \mathcal{Z}(xy - z^2, xz - y, x^2 - z)$ in $\mathbb{C}^3$. Using u, v as coordinates in $\mathbb{C}^2$, we find the reduced Gröbner basis G for the ideal $(u - x, v - y, xy - z^2, xz - y, x^2 - z)$ for the ordering $x > y > z > u > v$ contains only the polynomial $u^3 - v$ in $\mathbb{C}[u, v]$. The smallest algebraic set containing $\pi(V)$ is then the cubic $v = u^3$.

Affine Varieties

We next consider the question of whether an algebraic set can be decomposed into smaller algebraic sets and the corresponding algebraic formulation in terms of its coordinate ring.

Definition. A nonempty affine algebraic set V is called *irreducible* if it cannot be written as $V = V_1 \cup V_2$, where V_1 and V_2 are proper algebraic sets in V. An irreducible affine algebraic set is called an affine *variety.*

Equivalently, an algebraic set (which is a closed set in the Zariski topology) is irreducible if it cannot be written as the union of two proper, closed subsets.

Proposition 17.

(1) The affine algebraic set V is irreducible if and only if $\mathcal{I}(V)$ is a prime ideal.

(2) Every nonempty affine algebraic set V may be written uniquely in the form

$$V = V_1 \cup V_2 \cup \cdots \cup V_q$$

where each V_i is irreducible, and $V_i \not\subseteq V_j$ for all $j \neq i$ (i.e., the decomposition is "minimal" or "irredundant").

Proof: Let $I = \mathcal{I}(V)$ and suppose first that $V = V_1 \cup V_2$ is reducible, where V_1 and V_2 are proper closed subsets. Since $V_1 \neq V$, there is some function f_1 that vanishes on V_1 but not on V, i.e., $f_1 \in \mathcal{I}(V_1) - I$. Similarly, there is a function $f_2 \in \mathcal{I}(V_2) - I$. Then $f_1 f_2$ vanishes on $V_1 \cup V_2 = V$, so $f_1 f_2 \in I$ which shows that I is not a prime ideal. Conversely, if I is not a prime ideal, there exists $f_1, f_2 \in k[\mathbb{A}^n]$ such that $f_1 f_2 \in I$ but neither f_1 nor f_2 belongs to I. Let $V_1 = \mathcal{Z}(f_1) \cap V$ and $V_2 = \mathcal{Z}(f_2) \cap V$. Since the intersection of closed sets is closed, V_1 and V_2 are algebraic sets. Since neither f_1 nor f_2 vanishes on V, both V_1 and V_2 are proper subsets of V. Because $f_1 f_2 \in I$, $V \subseteq \mathcal{Z}(f_1 f_2) = \mathcal{Z}(f_1) \cup \mathcal{Z}(f_2)$, and so V is reducible. This proves (1).

To prove (2), let $\mathcal{S}$ be the collection of nonempty algebraic sets that cannot be written as a finite union of irreducible algebraic sets, and suppose by way of contradiction that $\mathcal{S} \neq \emptyset$. Let I_0 be a maximal element of the corresponding set of ideals, $\{\mathcal{I}(V) \mid V \in \mathcal{S}\}$, which exists (by Theorem 2) since $k[\mathbb{A}^n]$ is Noetherian. Then $V_0 = \mathcal{Z}(I_0)$ is a *minimal* element of $\mathcal{S}$. Since $V_0 \in \mathcal{S}$, it cannot be irreducible by the definition of $\mathcal{S}$. On the other hand, if $V_0 = V_1 \cup V_2$ for some proper, closed subsets V_1, V_2 of V_0, then by the minimality of V_0 both V_1 and V_2 may be written as finite unions of irreducible algebraic sets. Then V_0 may be written as a finite union of irreducible algebraic sets, a contradiction. This proves $\mathcal{S} = \emptyset$, i.e., every affine algebraic set has a decomposition into affine varieties.

To prove uniqueness, suppose V has two decompositions into affine varieties (where redundant terms have been removed from each decomposition):

$$V = V_1 \cup V_2 \cup \cdots \cup V_r = U_1 \cup U_2 \cup \cdots \cup U_s.$$

Then V_1 is contained in the union of the U_i. Since $V_1 \cap U_i$ is an algebraic set for each i, we obtain a decomposition of V_1 into algebraic subsets:

$$V_1 = (V_1 \cap U_1) \cup (V_1 \cap U_2) \cup \cdots \cup (V_1 \cap U_s).$$

Since V_1 is irreducible, we must have $V_1 = V_1 \cap U_j$ for some j, i.e., $V_1 \subseteq U_j$. By the symmetric argument we have $U_j \subseteq V_{j'}$ for some j'. Thus $V_1 \subseteq V_{j'}$, so $j' = 1$ and $V_1 = U_j$. Applying a similar argument for each V_i it follows that $r = s$ and that $\{V_1, \dots, V_r\} = \{U_1, \dots, U_s\}$. This completes the proof.

Corollary 18. An affine algebraic set V is a variety if and only if its coordinate ring $k[V]$ is an integral domain.

Proof: This follows immediately since $\mathcal{I}(V)$ is a prime ideal if and only if the quotient $k[V] = k[\mathbb{A}^n]/\mathcal{I}(V)$ is an integral domain (Proposition 13 of Chapter 7).

Definition. If V is a variety, then the field of fractions of the integral domain $k[V]$ is called the field of *rational functions* on V and is denoted by $k(V)$. The *dimension* of a variety V, denoted dim V, is defined to be the transcendence degree of $k(V)$ over k.

Examples

(1) Single points in $\mathbb{A}^n$ are affine varieties since their corresponding ideals in $k[\mathbb{A}^n]$ are maximal ideals. The coordinate ring of a point is isomorphic to k, which is also the field of rational functions. The dimension of a single point is 0. Any finite set is the union of its single point subsets, and this is its unique decomposition into affine subvarieties.

(2) The x-axis in $\mathbb{R}^2$ is irreducible since it has coordinate ring $\mathbb{R}[x, y]/(y) \cong \mathbb{R}[x]$, which is an integral domain. Similarly, the y-axis and, more generally, lines in $\mathbb{R}^2$ are also irreducible (cf. Exercise 23 in Section 1). Linear sets in $\mathbb{R}^n$ are affine varieties. The field of rational functions on the x-axis is the quotient field $\mathbb{R}(x)$ of $\mathbb{R}[x]$, which is why $\mathbb{R}(x)$ is called a rational function field. The dimension of the x-axis (or, more generally, any line) is 1.

(3) The union of the x and y axes in $\mathbb{R}^2$, namely $\mathcal{Z}(xy)$, is not a variety: $\mathcal{Z}(xy) = \mathcal{Z}(x) \cup \mathcal{Z}(y)$ is its unique decomposition into subvarieties. The corresponding coordinate ring $\mathbb{R}[x, y]/(xy)$ contains zero divisors.

(4) The hyperbola $xy = 1$ in $\mathbb{R}^2$ is a variety since we saw in Section 1 that its coordinate ring is the integral domain $\mathbb{R}[x, 1/x]$. Note that the two disjoint branches of the hyperbola (defined by $x > 0$ and $x < 0$) are not subvarieties (cf. also Exercises 12–13).

(5) If $V = \mathcal{Z}(l_1, l_2, \dots, l_m)$ is the zero set of *linear* polynomials $l_1, \dots, l_m$ in $k[x_1, \dots, x_m]$ and $V \neq \emptyset$, then V is an affine variety (called a *linear variety*). Note that determining whether $V \neq \emptyset$ is a linear algebra problem.

We end this section with some general ring-theoretic results that were originally motivated by their connection with decomposition questions in geometry.

Primary Decomposition of Ideals in Noetherian Rings

The second statement in Proposition 17 shows that any ideal of the form $\mathcal{I}(V)$ in $k[\mathbb{A}^n]$ may be written uniquely as a finite intersection of prime ideals, and by Hilbert's Nullstellensatz this applies in particular to all radical ideals when k is algebraically closed. In a large class of commutative rings (including all Noetherian rings) every ideal has a *primary decomposition*, which is a similar decomposition but allows ideals that are analogous to "prime powers" (but see the examples below). This decomposition can be considered as a generalization of the factorization of an integer $n \in \mathbb{Z}$ into the product of prime powers. We shall be primarily concerned with the case of Noetherian rings.

Definition. A proper ideal Q in the commutative ring R is called *primary* if whenever $ab \in Q$ and $a \notin Q$, then $b^n \in Q$ for some positive integer n. Equivalently, if $ab \in Q$ and $a \notin Q$, then $b \in \text{rad } Q$.

Some of the basic properties of primary ideals are given in the following proposition.

Proposition 19. Let R be a commutative ring with 1.

(1) Prime ideals are primary.

(2) The ideal Q is primary if and only if every zero divisor in R/Q is nilpotent.

(3) If Q is primary then rad Q is a prime ideal, and is the unique smallest prime ideal containing Q.

(4) If Q is an ideal whose radical is a maximal ideal, then Q is a primary ideal.

(5) Suppose M is a maximal ideal and Q is an ideal with $M^n \subseteq Q \subseteq M$ for some $n \geq 1$. Then Q is a primary ideal with rad $Q = M$.

Proof: The first two statements are immediate from the definition of a primary ideal. For (3), suppose $ab \in \text{rad } Q$. Then $a^m b^m = (ab)^m \in Q$, and since Q is primary, either $a^m \in Q$, in which case $a \in \text{rad } Q$, or $(b^m)^n \in Q$ for some positive integer n, in which case $b \in \text{rad } Q$. This proves that rad Q is a prime ideal, and it follows that rad Q is the smallest prime ideal containing Q (Proposition 12).

To prove (4) we pass to the quotient ring R/Q; by (2), it suffices to show that every zero divisor in this quotient ring is nilpotent. We are reduced to the situation where $Q = (0)$ and $M = \text{rad } Q = \text{rad}(0)$, which is the nilradical, is a maximal ideal. Since the nilradical is contained in every prime ideal (Proposition 12), it follows that M is the unique prime ideal, so also the unique maximal ideal. If d were a zero divisor, then the ideal (d) would be a proper ideal, hence contained in a maximal ideal. This implies that $d \in M$, hence every zero divisor is indeed nilpotent.

Finally, suppose $M^n \subseteq Q \subseteq M$ for some $n \geq 1$ where M is a maximal ideal. Then $Q \subseteq M$ so rad $Q \subseteq \text{rad } M = M$. Conversely, $M^n \subseteq Q$ shows that $M \subseteq \text{rad } Q$, so rad $Q = M$ is a maximal ideal, and Q is primary by (4).

Definition. If Q is a primary ideal, then the prime ideal $P = \text{rad } Q$ is called the *associated prime* to Q, and Q is said to *belong* to P (or to be *P-primary*).

It is easy to check that a finite intersection of P-primary ideals is again a P-primary ideal (cf. the exercises).

Examples

(1) The primary ideals in $\mathbb{Z}$ are 0 and the ideals (p^m) for p a prime and $m \geq 1$.

(2) For any field k, the ideal (x) in $k[x, y]$ is primary since it is a prime ideal. For any $n \geq 1$, the ideal $(x, y)^n$ is primary since it is a power of the maximal ideal (x, y).

(3) The ideal $Q = (x^2, y)$ in the polynomial ring $k[x, y]$ is primary since we have $(x, y)^2 \subseteq (x^2, y) \subseteq (x, y)$. Similarly, $Q' = (4, x)$ in $\mathbb{Z}[x]$ is a $(2, x)$-primary ideal.

(4) Primary ideals need not be powers of prime ideals. For example, the primary ideal Q in the previous example is not the power of a prime ideal, as follows. If $(x^2, y) = P^k$ for some prime ideal P and some $k \geq 1$, then $x^2, y \in P^k \subseteq P$ so $x, y \in P$. Then $P = (x, y)$, and since $y \notin (x, y)^2$, it would follow that $k = 1$ and $Q = (x, y)$. Since $x \notin (x^2, y)$, this is impossible.

(5) If R is Noetherian, and Q is a primary ideal belonging to the prime ideal P, then

$$P^m \subseteq Q \subseteq P$$

for some $m \geq 1$ by Proposition 14. If P is a maximal ideal, then the last statement in Proposition 19 shows that the converse also holds. This is not necessarily true if P

is a prime ideal that is *not maximal.* For example, consider the ideal $I = (x^2, xy)$ in $k[x, y]$. Then $(x^2) \subset I \subset (x)$, and (x) is a prime ideal, but I is not primary: $xy \in I$ and $x \notin I$, but no positive power of y is an element of I. This example also shows that an ideal whose radical is prime (but not maximal as in (4) of the proposition) is not necessarily primary.

(6) Powers of prime ideals need not be primary. For example, consider the quotient ring $R = \mathbb{R}[x, y, z]/(xy - z^2)$, the coordinate ring of the cone $z^2 = xy$ in $\mathbb{R}^3$, and let $P = (\bar{x}, \bar{z})$ be the ideal generated by $\bar{x}$ and $\bar{z}$ in R. This is a prime ideal in R since the quotient is $R/(\bar{x}, \bar{z}) \cong \mathbb{R}[x, y, z]/(x, z) \cong \mathbb{R}[y]$ (because $(xy - z^2) \subset (x, z)$). The ideal

$$P^2 = (\bar{x}^2, \bar{x}\bar{z}, \bar{z}^2) = (\bar{x}^2, \bar{x}\bar{z}, \bar{x}\bar{y}) = \bar{x}(\bar{x}, \bar{y}, \bar{z}),$$

however, is not primary: $\bar{x}\bar{y} = \bar{z}^2 \in P^2$, but $\bar{x} \notin P^2$, and no power of $\bar{y}$ is in P^2. Note that P^2 is another example of an ideal that is not primary whose radical is prime.

(7) Suppose R is a U.F.D. If π is an irreducible element of R then it is easy to see that the powers (π^n) for $n = 1, 2, \ldots$ are (π)-primary ideals. Conversely, suppose Q is a (π)-primary ideal, and let n be the largest integer with $Q \subseteq (\pi^n)$ (such an integer exists since, for example, $\pi^k \in Q$ for some $k \geq 1$, so $n \leq k$). If q is an element of Q not contained in (π^{n+1}), then $q = r\pi^n$ for some $r \in R$ and $r \notin (\pi)$. Since $r \notin (\pi)$ and Q is (π)-primary, it follows that $\pi^n \in Q$. This shows that $Q = (\pi^n)$.

In the examples above, the ideal (x^2, xy) in $k[x, y]$ is not a primary ideal, but it can be written as the intersection of primary ideals: $(x^2, xy) = (x) \cap (x, y)^2$.

Definition.

(1) An ideal I in R has a *primary decomposition* if it may be written as a finite intersection of primary ideals:

$$I = \bigcap_{i=1}^{m} Q_i \qquad Q_i \text{ a primary ideal.}$$

(2) The primary decomposition above is *minimal* and the Q_i are called the *primary components of I* if

(a) no primary ideal contains the intersection of the remaining primary ideals, i.e., $Q_i \not\supseteq \cap_{j \neq i} Q_j$ for all i, and

(b) the associated prime ideals are all distinct: rad $Q_i \neq$ rad Q_j for $i \neq j$.

We now prove that in a Noetherian ring every proper ideal has a minimal primary decomposition. This result is often called the Lasker–Noether Decomposition Theorem, since it was first proved for polynomial rings by the chess master Emanuel Lasker and the proof was later greatly simplified and generalized by Emmy Noether.

Definition. A proper ideal I in the commutative ring R is said to be *irreducible* if I cannot be written nontrivially as the intersection of two other ideals, i.e., if $I = J \cap K$ with ideals J, K implies that $I = J$ or $I = K$.

It is easy to see that a prime ideal is irreducible (see Exercise 11 in Section 7.4). The ideal $(x, y)^2$ in $k[x, y]$ in Example 2 earlier shows that primary ideals need not be irreducible since it is the intersection of the ideals $(x) + (x, y)^2 = (x, y^2)$ and $(y)+(x, y)^2 = (y, x^2)$. In a Noetherian ring, however, irreducible ideals are necessarily primary:

Proposition 20. Let R be a Noetherian ring. Then

(1) every irreducible ideal is primary, and

(2) every proper ideal in R is a finite intersection of irreducible ideals.

Proof: To prove (1) let Q be an irreducible ideal and suppose that $ab \in Q$ and $b \notin Q$. It is easy to check that for any fixed n the set of elements $x \in R$ with $a^n x \in Q$ is an ideal, A_n, in R. Clearly $A_1 \subseteq A_2 \subseteq \dots$ and since R is Noetherian this ascending chain of ideals must stabilize, i.e., $A_n = A_{n+1} = \dots$ for some $n > 0$. Consider the two ideals $I = (a^n) + Q$ and $J = (b) + Q$ of R, each containing Q. If $y \in I \cap J$ then $y = a^n z + q$ for some $z \in R$ and $q \in Q$. Since $ab \in Q$, it follows that $aJ \subseteq Q$, and in particular $ay \in Q$. Then $a^{n+1}z = ay - aq \in Q$, so $z \in A_{n+1} = A_n$. But $z \in A_n$ means that $a^n z \in Q$, so $y \in Q$. It follows that $I \cap J = Q$. Since Q is irreducible and $(b) + Q \neq Q$ (since $b \notin Q$), we must have $a^n \in Q$, which shows that Q is primary.

The proof of (2) is the same as the proof of the second statement in Proposition 17. Let $\mathcal{S}$ be the collection of ideals of R that cannot be written as a finite intersection of irreducible ideals. If $\mathcal{S}$ is not empty, then since R is Noetherian, there is a maximal element I in $\mathcal{S}$. Then I is not itself irreducible, so $I = J \cap K$ for some ideals J and K distinct from I. Then $I \subset J$ and $I \subset K$ and the maximality of I implies that neither J nor K is in $\mathcal{S}$. But this means that both J and K can be written as finite intersections of irreducible ideals, hence the same would be true for I. This is a contradiction, so $\mathcal{S} = \emptyset$, which completes the proof of the proposition.

It is immediate from the previous proposition that in a Noetherian ring every proper ideal has a primary decomposition. If any of the primary ideals in this decomposition contains the intersection of the remaining primary ideals, then we may simply remove this ideal since this will not change the intersection. Hence we may assume the decomposition satisfies (a) in the definition of a minimal decomposition. Since a finite intersection of P-primary ideals is again P-primary (Exercise 31), replacing the primary ideals in the decomposition with the intersections of all those primary ideals belonging to the same prime, we may also assume the decomposition satisfies (b) in the definition of a minimal decomposition. This proves the first statement of the following:

Theorem 21. *(Primary Decomposition Theorem)* Let R be a Noetherian ring. Then every proper ideal I in R has a minimal primary decomposition. If

$$I = \bigcap_{i=1}^{m} Q_i = \bigcap_{i=1}^{n} Q_i'$$

are two minimal primary decompositions for I then the sets of associated primes in the two decompositions are the same:

$$\{\operatorname{rad} Q_1, \operatorname{rad} Q_2, \dots, \operatorname{rad} Q_m\} = \{\operatorname{rad} Q_1', \operatorname{rad} Q_2', \dots, \operatorname{rad} Q_n'\}.$$

Moreover, the primary components Q_i belonging to the minimal elements in this set of associated primes are uniquely determined by I.

Proof: The proof of the uniqueness of the set of associated primes is outlined in the exercises, and the proof of the uniqueness of the primary components associated to the minimal primes will be given in Section 4.

Definition. If I is an ideal in the Noetherian ring R then the associated prime ideals in any primary decomposition of I are called the *associated prime ideals of* I. If an associated prime ideal P of I does not contain any other associated prime ideal of I then P is called an *isolated prime ideal*; the remaining associated prime ideals of I are called *embedded prime ideals.*

The prime ideals associated to an ideal I provide a great deal of information about the ideal I (cf. for example Exercises 41 and 43):

Corollary 22. Let I be a proper ideal in the Noetherian ring R.

(1) A prime ideal P contains the ideal I if and only if P contains one of the associated primes of I, hence if and only if P contains one of the isolated primes of I, i.e., the isolated primes of I are precisely the minimal elements in the set of all prime ideals containing I. In particular, there are only finitely many minimal elements among the prime ideals containing I.

(2) The radical of I is the intersection of the associated primes of I, hence also the intersection of the isolated primes of I.

(3) There are prime ideals $P_1, \dots, P_n$ (not necessarily distinct) containing I such that $P_1 P_2 \cdots P_n \subseteq I$.

Proof: The first statement in (1) is an exercise (cf. Exercise 37), and the remainder of (1) follows. Then (2) follows from (1) and Proposition 12, and (3) follows from (2) and Proposition 14.

The last statement in Theorem 21 states that not only the isolated primes, but also the primary components belonging to the isolated primes, are uniquely determined by I. In general the primary decomposition of an ideal I is itself not unique.

Examples

(1) Let $I = (x^2, xy)$ in $\mathbb{R}[x, y]$. Then

$$(x^2, xy) = (x) \cap (x, y)^2 = (x) \cap (x^2, y)$$

are two minimal primary decompositions for I. The associated primes for I are (x) and $\text{rad}((x, y)^2) = \text{rad}((x^2, y)) = (x, y)$. The prime (x) is the only isolated prime since $(x) \subset (x, y)$, and (x, y) is an embedded prime. A prime ideal P contains I if and only if P contains (x). The (x)-primary component of I corresponding to this isolated prime is just (x) and occurs in both primary decompositions; the (x, y)-primary component of I corresponding to this embedded prime is not uniquely determined — it is $(x, y)^2$ in the first decomposition and is (x^2, y) in the second. The radical of I is the isolated prime (x).

This example illustrates the origin of the terminology: in general the irreducible components of the algebraic space $\mathcal{Z}(I)$ defined by I are the zero sets of the isolated primes for I, and the zero sets of the embedded primes are irreducible subspaces of these components (so are "embedded" in the irreducible components). In this example, $\mathcal{Z}(I)$ is the set of points with $x^2 = xy = 0$, which is just the y-axis in $\mathbb{R}^2$. There is only one irreducible component of this algebraic space (namely the y-axis), which is the locus for the isolated prime (x). The locus for the embedded prime (x, y) is the origin $(0, 0)$, which is an irreducible subspace embedded in the y-axis.

(2) Suppose R is a U.F.D. If $a = p_1{}^{e_1} \cdots p_m{}^{e_m}$ is the unique factorization into distinct prime powers of the element $a \in R$, then $(a) = (p_1)^{e_1} \cap \cdots \cap (p_m)^{e_m}$ is the minimal primary decomposition of the principal ideal (a). The associated primes to (a) are $(p_1), \ldots, (p_m)$ and are all isolated. The primary decomposition of ideals is a generalization of the factorization of elements into prime powers. See also Exercise 44 for a characterization of U.F.D.s in terms of minimal primary decompositions.

For any Noetherian ring, an ideal I is radical if and only if the primary components of a minimal primary decomposition of I are all *prime* ideals (in which case this primary decomposition is unique), cf. Exercise 43. This generalizes the observation made previously that Proposition 17 together with Hilbert's Nullstellensatz shows that any radical ideal in $k[\mathbb{A}^n]$ may be written uniquely as a finite intersection of prime ideals when the field k is algebraically closed — this is the algebraic statement that an algebraic set can be decomposed uniquely into the union of irreducible algebraic sets.

EXERCISES

1. Prove (3) of Corollary 22 directly by considering the collection $\mathcal{S}$ of ideals that do not contain a finite product of prime ideals. [If I is a maximal element in $\mathcal{S}$, show that since I is not prime there are ideals J, K properly containing I (hence not in $\mathcal{S}$) with $JK \subseteq I$.]

2. Let I and J be ideals in the ring R. Prove the following statements:
(a) If $I^k \subseteq J$ for some $k \geq 1$ then $\operatorname{rad} I \subseteq \operatorname{rad} J$.
(b) If $I^k \subseteq J \subseteq I$ for some $k \geq 1$ then $\operatorname{rad} I = \operatorname{rad} J$.
(c) $\operatorname{rad}(IJ) = \operatorname{rad}(I \cap J) = \operatorname{rad} I \cap \operatorname{rad} J$.
(d) $\operatorname{rad}(\operatorname{rad} I) = \operatorname{rad} I$.
(e) $\operatorname{rad} I + \operatorname{rad} J \subseteq \operatorname{rad}(I + J)$ and $\operatorname{rad}(I + J) = \operatorname{rad}(\operatorname{rad} I + \operatorname{rad} J)$.

3. Prove that the intersection of two radical ideals is again a radical ideal.

4. Let $I = \mathfrak{m}_1\mathfrak{m}_2$ be the product of the ideals $\mathfrak{m}_1 = (x, y)$ and $\mathfrak{m}_2 = (x-1, y-1)$ in $\mathbb{F}_2[x, y]$. Prove that I is a radical ideal. Prove that the ideal $(x^3 - y^2)$ is a radical ideal in $\mathbb{F}_2[x, y]$.

5. If $I = (xy, (x - y)z) \subset k[x, y, z]$ prove that $\operatorname{rad} I = (xy, xz, yz)$. For this ideal prove directly that $\mathcal{Z}(I) = \mathcal{Z}(\operatorname{rad} I)$, that $\mathcal{Z}(I)$ is not irreducible, and that $\operatorname{rad} I$ is not prime.

6. Give an example to show that over a field k that is not algebraically closed the containment $I \subseteq \mathcal{I}(\mathcal{Z}(I))$ can be proper even when I is a radical ideal.

7. Suppose R and S are rings and $\varphi : R \to S$ is a ring homomorphism. If I is an ideal of R show that $\varphi(\operatorname{rad} I) \subseteq \operatorname{rad}(\varphi(I))$. If in addition φ is surjective and I contains the kernel of φ show that $\varphi(\operatorname{rad} I) = \operatorname{rad}(\varphi(I))$.

8. Suppose the prime ideal P contains the ideal I. Prove that P contains the radical of I.

9. Prove that for any field k the map $\mathcal{Z}$ in the Nullstellensatz is always surjective and the map $\mathcal{I}$ in the Nullstellensatz is always injective. [Use property (10) of the maps $\mathcal{Z}$ and $\mathcal{I}$ in Section 1.] Give examples (over a field k that is not algebraically closed) where $\mathcal{Z}$ is not injective and $\mathcal{I}$ is not surjective.

10. Prove that for k a finite field the Zariski topology is the same as the discrete topology: every subset is closed (and open).

11. Let V be a variety in $\mathbb{A}^n$ and let U_1 and U_2 be two subsets of $\mathbb{A}^n$ that are open in the Zariski topology. Prove that if $V \cap U_1 \neq \emptyset$ and $V \cap U_2 \neq \emptyset$ then $V \cap U_1 \cap U_2 \neq \emptyset$. Conclude that *any* nonempty open subset of a variety is *everywhere dense* in the Zariski topology (i.e., its closure is all of V).

12. Use the fact that nonempty open sets of an affine variety are everywhere dense to prove that an affine variety is connected in the Zariski topology. (A topological space is *connected* if it is not the union of two disjoint, proper, open subsets.)

13. Prove that the affine algebraic set V is connected in the Zariski topology if and only if $k[V]$ is not a direct sum of two nonzero ideals. Deduce from this that a variety is connected in the Zariski topology.

14. Prove that if k is an infinite field, then the varieties in $\mathbb{A}^1$ are the empty set, the whole space, and the one point subsets. What are the varieties in $\mathbb{A}^1$ in the case of a finite field k?

15. Suppose V is a hypersurface in $\mathbb{A}^n$ and $\mathcal{I}(V) = (f)$ for some nonconstant polynomial $f \in k[x_1, x_2, \dots, x_n]$. Prove that V is a variety if and only if f is irreducible.

16. Suppose $V \subseteq \mathbb{A}^n$ is an affine variety and $f \in k[V]$. Prove that the *graph* of f (cf. Exercise 25 in Section 1) is an affine variety.

17. Prove that any permutation of the elements of a field k is a continuous map from $\mathbb{A}^1$ to itself in the Zariski topology on $\mathbb{A}^1$. Deduce that if k is an infinite field, there are Zariski continuous maps from $\mathbb{A}^1$ to itself that are not polynomials.

18. Let V be an affine algebraic set in $\mathbb{A}^n$ over $k = \mathbb{C}$.

(a) Prove that morphisms of algebraic sets over $\mathbb{C}$ are continuous in the Euclidean topology (the topology on $\mathbb{C}^n$ obtained by identifying $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ with its usual Euclidean topology).

(b) Prove that V is a closed set in the Euclidean topology on $\mathbb{C}^n$ (so the Zariski closed sets of $\mathbb{A}^n$ over $\mathbb{C}$ are also Euclidean closed).

(c) Give an example of a set that is closed in the Euclidean topology but is not closed in the Zariski topology, i.e., is not an affine algebraic set (so the Euclidean topology is "finer" than the Zariski topology).

19. Give an example of an injective k-algebra homomorphism $\widetilde{\varphi} : k[W] \to k[V]$ whose associated morphism $\varphi : V \to W$ is not surjective.

20. Suppose $\varphi : V \to W$ is a surjective morphism of affine algebraic sets. Prove that if V is a variety then W is a variety.

21. Let V be an algebraic set in $\mathbb{A}^n$ and let $f \in k[V]$. Define $V_f = \{v \in V \mid f(v) \neq 0\}$.

(a) Show that V_f is a Zariski open set in V (called a *principal open set* in V).

(b) Let J be the ideal in $k[x_1, \dots, x_n, x_{n+1}]$ generated by $\mathcal{I}(V)$ and $x_{n+1}f - 1$, and let $W = \mathcal{Z}(J) \subseteq \mathbb{A}^{n+1}$. Show that $J = \mathcal{I}(W)$ and that the map $\pi : \mathbb{A}^{n+1} \to \mathbb{A}^n$ by projection onto the first n coordinates is a Zariski continuous bijection from W onto V_f (so the principal *open* set V_f in V may be embedded as a *closed* set in some (larger) affine space).

(c) If U is any open set in V show that $U = V_{f_1} \cup \cdots \cup V_{f_m}$ for some $f_1, \dots, f_m \in k[V]$. (This shows that the principal open sets form a *base* for the Zariski topology.)

22. Prove that $GL_n(k)$ is an open affine algebraic set in $\mathbb{A}^{n^2}$ and can be embedded as a closed affine algebraic set in $\mathbb{A}^{n^2+1}$. In particular, deduce that the set $k^\times$ of nonzero elements in

$\mathbb{A}^1$ embeds into $\mathbb{A}^2$ as the hyperbola $xy = 1$. [Use the preceding exercise.]

23. Show that if k is infinite then $\{(a, a^2, a^3) \mid a \in k\} \subset \mathbb{A}^3$ is an affine algebraic variety. If k is finite show that this set is always reducible.

24. Let $V = \mathcal{Z}(xz - y^2, yz - x^3, z^2 - x^2y) \subset \mathbb{A}^3$. Show that if k is infinite then V is an affine variety. [Use Exercise 26 of Section 1 and Exercise 20.]

25. Suppose $f(x) = x^3 + ax^2 + bx + c$ is an irreducible cubic in $\mathbb{Q}[x]$ of discriminant D. Let $I = (x + y + z + a, xy + xz + yz - b, xyz + c)$ in $\mathbb{Q}[x, y, z]$.

(a) Prove that I is a prime ideal if and only if D is not a square in $\mathbb{Q}$, in which case I is a maximal ideal and $\mathbb{Q}[x, y, z]/I$ is a splitting field for $f(x)$ over $\mathbb{Q}$.

(b) If $D = r^2$, prove that the primary decomposition of I is $I = Q_+ \cap Q_-$ where $Q_\pm = (I, (x - y)(x - z)(y - z) \pm r)$. Prove Q_+ and Q_- are maximal ideals, and $\mathbb{Q}[x, y, z]$ modulo Q_+ or Q_- is a splitting field for $f(x)$ over $\mathbb{Q}$.

26. A topological space X is called *quasicompact* if whenever any collection of closed subsets V_i of X has empty intersection, then some finite number of these also has empty intersection, i.e.,

$$\text{whenever } \bigcap_i V_i = \emptyset \text{ there exists } V_{i_1}, V_{i_2}, \ldots, V_{i_N} \text{ such that } \bigcap_{t=1}^{N} V_{i_t} = \emptyset.$$

Prove that every affine algebraic set is quasicompact. [Translate the definition into a property of ideals in $k[x_1, \ldots, x_n]$.] (A quasicompact and Hausdorff space is called *compact*.)

27. When k is an infinite field prove that the Zariski topology on k^2 is not the same as taking the Zariski topology on k and then forming the product topology on $k \times k$. [By Exercise 14 of Section 1, in the product topology on $k \times k$ the Zariski closed sets in $k \times k$ are finite unions of sets of the form $\{a\} \times \{b\}$, $\{a\} \times k$ and $k \times \{b\}$, for any $a, b \in k$.]

28. Prove that each of the following rings have infinitely many minimal prime ideals, and that (0) is not the intersection of any finite number of these (so (0) does not have a primary decomposition in these rings):

(a) the infinite direct product ring $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \times \cdots$ (which is a Boolean ring, cf. Exercise 23 in Section 7.4).

(b) $k[x_1, x_2, \ldots]/(x_1x_2, x_3x_4, \ldots, x_{2i-1}x_{2i}, \ldots)$, where $x_1, x_2, \ldots$ are independent variables over the field k.

29. Suppose that A and B are ideals with $AB \subseteq Q$ for a primary ideal Q. Prove that if $A \not\subseteq Q$ then $B \subset \operatorname{rad} Q$.

30. Let Q be a P-primary ideal and suppose A is an ideal not contained in Q. Define $A' = \{r \in R \mid rA \subseteq Q\}$ to be the elements of R that when multiplied by elements of A give elements of Q. Prove that A' is a P-primary ideal.

31. Prove that if Q_1 and Q_2 are primary ideals belonging to the same prime ideal P, then $Q_1 \cap Q_2$ is a primary ideal belonging to P. Conclude that a finite intersection of P-primary ideals is again P-primary.

32. Prove that if Q_1 and Q_2 are primary ideals belonging to the same *maximal* ideal M, then $Q_1 + Q_2$ and Q_1Q_2 are primary ideals belonging to M. Conclude that finite sums and finite products of M-primary ideals are again M-primary.

33. Let $I = (x^2, xy, xz, yz)$ in $k[x, y, z]$. Prove that a primary decomposition of I is $I = (x, y) \cap (x, z) \cap (x, y, z)^2$, determine the isolated and embedded primes of I, and find $\operatorname{rad} I$.

34. Suppose $\varphi : R \to S$ is a surjective ring homomorphism. Prove that an ideal Q in R containing the kernel of φ is primary if and only if $\varphi(Q)$ is primary in S, and when this is the case the prime associated to $\varphi(Q)$ is the image $\varphi(P)$ of the prime P associated to Q.

35. Suppose $\varphi : R \to S$ is a ring homomorphism.

(a) Suppose I is an ideal of R containing $\ker \varphi$ with minimal primary decomposition $I = Q_1 \cap \cdots \cap Q_m$ with rad $Q_i = P_i$. If φ is a surjective homomorphism prove that $\varphi(I) = \varphi(Q_1) \cap \cdots \cap \varphi(Q_m)$, where rad $\varphi(Q_i)$ is given by $\varphi(P_i)$, is a minimal primary decomposition of $\varphi(I)$. [Use the previous exercise.]

(b) Suppose I is an ideal of S with minimal primary decomposition $I = Q_1 \cap \cdots \cap Q_m$ withrad $Q_i = P_i$. Prove that $\varphi^{-1}(I) = \varphi^{-1}(Q_1) \cap \cdots \cap \varphi^{-1}(Q_m)$, where rad $\varphi^{-1}(Q_i)$ is given by $\varphi^{-1}(P_i)$, is a primary decomposition of $\varphi^{-1}(I)$, and is minimal if φ is surjective.

36. Let $I = (xy, x - yz)$ in $k[x, y, z]$. Prove that $(x, z) \cap (y^2, x - yz)$ is a minimal primary decomposition of I. [Consider the ring homomorphism $\varphi : k[x, y, z] \to k[y, z]$ given by mapping x to yz, y to y, and z to z and use the previous exercise.]

37. Prove that a prime ideal P contains the ideal I if and only if P contains one of the associated primes of a minimal primary decomposition of I. [Use Exercise 3 and Exercise 11 in Section 7.4.]

38. Show that every associated prime ideal for a radical ideal is isolated. [Suppose that $P_2 = \text{rad } Q_2 \subseteq P_1 = \text{rad } Q_1$ in the decomposition of Theorem 21 for the radical ideal I. Show that if $a \in Q_2 \cap \cdots \cap Q_m \subseteq P_2$ then $a^n \in I$ for some $n \geq 1$, conclude that $a \in Q_1$ and derive a contradiction to the minimality of the primary decomposition.]

39. Fix an element a in the ring R. For any ideal I in the ring R let $I_a = \{r \in R \mid ar \in I\}$.

(a) Prove that I_a is an ideal and $I_a = R$ if and only if $a \in I$.

(b) Prove that $(I \cap J)_a = I_a \cap J_a$ for ideals I and J.

(c) Suppose that Q is a P-primary ideal and that $a \notin Q$. Prove that Q_a is a P-primary ideal and that $Q_a = Q$ if $a \notin P$.

40. With notation as in the previous exercise, suppose $I = Q_1 \cap \cdots \cap Q_m$ is a minimal primary decomposition of the ideal I and let P_i be the prime ideal associated to Q_i.

(a) Prove that $I_a = (Q_1)_a \cap \cdots \cap (Q_m)_a$ and that $\text{rad}(I_a) = \text{rad}((Q_1)_a) \cap \cdots \cap \text{rad}((Q_m)_a)$.

(b) Prove that $\text{rad}(I_a)$ is the intersection of the prime ideals P_i for which $a \notin Q_i$. [Use the previous exercise.]

(c) Prove that if $\text{rad}(I_a)$ is a prime ideal then $\text{rad}(I_a) = P_j$ for some j. [Use the fact that prime ideals are irreducible.]

(d) For each $i = 1, \ldots, m$, prove that $\text{rad}(I_a) = P_i$ for some $a \in R$. [Show there exists an $a \in R$ with $a \notin Q_i$ but $a \in Q_j$ for all $j \neq i$.]

(e) Show from (c) and (d) that the associated primes for a minimal primary decomposition are precisely the collection of prime ideals among the ideals $\text{rad}(I_a)$ for $a \in R$, and conclude that they are uniquely determined by I independent of the minimal primary decomposition.

41. Let $P_1, \ldots, P_m$ be the associated prime ideals of the ideal (0) in the Noetherian ring R.

(a) Show that $P_1 \cap \cdots \cap P_m$ is the collection of nilpotent elements in R. [Apply Corollary 22 to $I = (0)$.]

(b) Show that $P_1 \cup \cdots \cup P_m$ is the collection of zero divisors in R. [Let $I = (0)$ in the previous exercise and show that the set of zero divisors is given by the set $\cup_{a \in R - \{0\}} (0)_a = \cup_{a \in R - \{0\}} \text{rad}((0)_a)$.]

42. Suppose R is a Noetherian ring. Prove that R is either an integral domain, has nonzero nilpotent elements, or has at least two minimal prime ideals. [Use the previous exercise.]

43. Prove that the ideal I in the Noetherian ring R is radical if and only if the primary components of a minimal primary decomposition are all prime ideals, and conclude that in this case the minimal primary decomposition is unique. [If $I = Q_1 \cap \cdots \cap Q_m$ is radical with Q_i a P_i-primary component of a minimal decomposition, show that if $a \in P_1 \cap \cdots \cap P_m$ then some power of a is in I, hence $a \in I$ since I is radical. Deduce that $I = P_1 \cap \cdots \cap P_m$ and show that this is also a minimal primary decomposition, i.e., for any i there exists b with $b \notin P_i$, but $b \in P_j$ for $j \neq i$. If $a \in P_i$, show that $ab \in Q_i$, and that $a \in Q_i$. Conclude that $Q_i = P_i$.]

44. Prove that a Noetherian integral domain R is a U.F.D. if and only if for every $a \in R$ the isolated primes associated to the principal ideal (a) are principal ideals. [See Example 2 following Corollary 22. To prove R is a U.F.D., show that an irreducible $a \in R$ is prime and then follow the proof of Theorem 14 in Section 8.3.]

45. Let R be the ring of all real valued functions on the open interval $(-1, 1)$ that have derivatives of all orders (the ring of C^∞ functions). Let

$$F(x) = \begin{cases} e^{-1/x^4} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

(you may assume $F \in R$ and $F^{(n)}(0) = 0$ for all $n \geq 0$). Let (F) be the principal ideal generated by F and let $A = \operatorname{rad}((F))$. Let M be the (maximal) ideal of all functions in R that are zero at $x = 0$ and let $P = \cap_{n=1}^{\infty} M^n$.

(a) Prove that $M = (x)$ is the ideal generated by the function x in R and that $M^n = (x^n)$ consists of the functions whose first $n - 1$ derivatives vanish at the origin.

(b) Prove that R is not Noetherian (compare Exercise 33 in Section 7.4). [One approach is the following: Let $G(x)$ be the function that is 0 for $x < 0$ and is equal to $F(x)$ for $x \geq 0$. Let I_n be the ideal of functions in R vanishing for all $x \leq 1/n$. Use translates of $G(x)$ to show that $I_1 \subset I_2 \subset I_3 \subset \cdots$ is an infinite ascending chain.]

(c) Prove that P consists of the functions all of whose derivatives are zero at $x = 0$ (i.e., the functions whose associated Taylor series at $x = 0$ is identically zero), and that P is a prime ideal.

(d) Prove that $F \in P$ and deduce that $A \subseteq P$.

(e) Prove that $A \neq P$. [Let $G(x) = e^{-1/x^2}$ when $x \neq 0$ and $G(0) = 0$. Show that $G \in P$ but $G \notin A$.]

(f) Show that there is a prime ideal Q containing (F) with $Q \neq P, M$. Prove that $Q \subset P$ i.e., there are nonzero prime ideals properly contained in P.

46. Let $\mathcal{A}$ be any ideal in $R = k[x_1, \ldots, x_n, y_1, \ldots, y_m]$.

(a) Show that $\operatorname{rad}(\mathcal{A} \cap k[y_1, \ldots, y_m]) = \operatorname{rad} \mathcal{A} \cap k[y_1, \ldots, y_m]$.

(b) Suppose $(f_1, \ldots, f_s)$ is an ideal in $k[x_1, \ldots, x_n]$. Let $F_1, \ldots, F_t$ be generators for the radical of $(f_1, \ldots, f_s)$, computed in $k[x_1, \ldots, x_n]$. Suppose J is an ideal in R and let $\mathcal{A} = J + (f_1, \ldots, f_s)$, $\mathcal{B} = J + (F_1, \ldots, F_t)$ as ideals in R. Prove that $\operatorname{rad} \mathcal{A} = \operatorname{rad} \mathcal{B}$.

(c) Conclude from (a) and (b) that $\mathcal{A} = (y_1 - x_1, \ldots, y_m - x_m, f_1, \ldots, f_s) \cap k[y_1, \ldots, y_m]$ and $\mathcal{B} = (y_1 - x_1, \ldots, y_m - x_m, F_1, \ldots, F_t) \cap k[y_1, \ldots, y_m]$ have the same zero sets over an algebraically closed field k. [Use Hilbert's Nullstellensatz.]

47. Determine the Zariski closure in $\mathbb{C}^3$ of the points on the curve $\{(a^2, a^3, a^4) \mid a \in \mathbb{C}\}$.

48. Show that $\mathcal{Z}(x^3 - xyz + z^2)$ is the smallest algebraic set in $\mathbb{R}^3$ containing the points $\{(st, s + t, s^2t) \mid s, t \in \mathbb{R}\}$.

49. Show that $\mathcal{Z}(x^3z^2 - 3xy^2z^2 - y^6 - z^4)$ is the smallest algebraic set in $\mathbb{R}^3$ containing the points $\{(s^2 + t^2, st, s^3) \mid s, t \in \mathbb{R}\}$.

50. Find equations defining the Zariski closure of the set of points $\{(s^4, s^3t, st^3, t^4) \mid s, t \in \mathbb{R}\}$.

51. Show that $V = \mathcal{Z}(x^2 - y^2z)$ (the *Whitney umbrella surface*) is the smallest algebraic set in $\mathbb{R}^3$ containing the points $S = \{(st, s, t^2) \mid s, t \in \mathbb{R}\}$. Show that S is not Zariski closed in V (the missing points explain the name for the surface). Do the same over $\mathbb{C}$, but show that in this case $S = V$ is closed.

52. Let $V = \mathcal{Z}(xz^2 - w^3, xw^2 - y^4, y^4z^2 - w^5) \subset \mathbb{C}^4$. Determine the Zariski closure of the image of V under the projection $\pi((x, y, z, w)) = (x, y, z)$.

53. Let $V = \mathcal{Z}(xy - 1)$ in $\mathbb{A}^2$ and let S be the projection of V onto the x-axis in $\mathbb{A}^1$.

(a) If $k = \mathbb{R}$, show that $\mathcal{I}(V) = (xy - 1) \subset \mathbb{R}[x, y]$ and that $(u - x, xy - 1) \cap \mathbb{R}[u] = 0$ in $\mathbb{R}[x, y, u]$. Use Propositions 8 and 16 to conclude that the Zariski closure of S is $\mathbb{A}^1$ and show that S is not itself closed.

(b) If $k = \mathbb{F}3$, show that $\mathcal{I}(V) = (xy - 1, x^3 - x, y^3 - y) \subset \mathbb{F}3[x, y]$ and that $(u - x, xy - 1, x^3 - x, y^3 - y) \cap \mathbb{F}3[u] = (u^2 - 1)$ in $\mathbb{F}3[x, y, u]$. Use Propositions 8 and 16 to conclude that S is Zariski closed in $\mathbb{A}^1$.

54. Recall the *ideal quotient* $(I : J) = \{r \in R \mid rJ \in I\}$ of two ideals I, J in a ring R (cf. Exercise 34 *ff.* in Section 9.6). Clearly $I \subseteq (I : J)$.

(a) Show that $\mathcal{Z}(I) - \mathcal{Z}(J)$, the set of elements of $\mathcal{Z}(I)$ not lying in $\mathcal{Z}(J)$, is contained in $\mathcal{Z}((I : J))$ and conclude that the Zariski closure of $\mathcal{Z}(I) - \mathcal{Z}(J)$ is contained in $\mathcal{Z}((I : J))$.

(b) Show that if k is algebraically closed and I is a radical ideal then $\mathcal{Z}((I : J))$ is precisely the Zariski closure of $\mathcal{Z}(I) - \mathcal{Z}(J)$.

(c) Show that if V and W are affine algebraic sets then $(\mathcal{I}(V) : \mathcal{I}(W)) = \mathcal{I}(V - W)$.

15.3 INTEGRAL EXTENSIONS AND HILBERT'S NULLSTELLENSATZ

In this section we consider the important concept of an integral extension of rings, which is a generalization to rings of algebraic extensions of fields. This leads to the definition of the "integers" in finite extensions of $\mathbb{Q}$ (the basic subject of the branch of mathematics called algebraic number theory) and is also related to the existence of tangent lines for algebraic curves.

Definition. Suppose R is a subring of the commutative ring S with $1 = 1_S \in R$.

(1) An element $s \in S$ is *integral over* R if s is the root of a monic polynomial in $R[x]$.

(2) The ring S is an *integral extension of* R or just *integral over* R if every $s \in S$ is integral over R.

(3) The *integral closure* of R in S is the set of elements of S that are integral over R.

(4) The ring R is said to be *integrally closed in* S if R is equal to its integral closure in S. The integral closure of an integral domain R in its field of fractions is called the *normalization of* R. An integral domain is called *integrally closed* or *normal* if it is integrally closed in its field of fractions.

Before giving some examples of integral extensions we prove some basic properties of integral elements analogous to those of algebraic elements over fields.

Proposition 23. Let R be a subring of the commutative ring S with $1 \in R$ and let $s \in S$. Then the following are equivalent:

(1) s is integral over R,

(2) $R[s]$ is a finitely generated R-module (where $R[s]$ is the ring of all R-linear combinations of powers of s), and

(3) $s \in T$ for some subring T, $R \subseteq T \subseteq S$, that is a finitely generated R-module.

Proof: Suppose first that (1) holds and let s be a root of the monic polynomial $x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in R[x]$. Then

$$s^n = -(a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \cdots + a_0)$$

and so s^n, and then all higher powers of s, can be expressed as R-linear combinations of $s^{n-1}, \ldots, s, 1$. Hence $R[s] = R1 + Rs + \cdots + Rs^{n-1}$ is finitely generated as an R-module, which gives (2).

If (2) holds, then (3) holds with $T = R[s]$.

Suppose that (3) holds and let $v_1, v_2, \ldots, v_n$ be a finite generating set for T. Then for $i = 1, 2, \ldots, n$ the element sv_i is an element of T since T is a ring, and so can be written as R-linear combinations of $v_1, \ldots, v_n$:

$$sv_i = \sum_{j=1}^{n} a_{ij}v_j,$$

i.e.,

$$0 = \sum_{j=1}^{n}(\delta_{ij}s - a_{ij})v_j \qquad i = 1, 2, \ldots, n$$

where δ_{ij} is the Kronecker delta. If B is the $n \times n$ matrix whose i, j entry is $\delta_{ij}s - a_{ij}$, and v is the $n \times 1$ column vector whose entries are $v_1, \ldots, v_n$, then these equations are simply $Bv = 0$. It follows from Cramer's Rule that $(\det B)v_i = 0$ for all i (cf. Exercise 3, Section 11.4). Since $1 \in T$ is an R-linear combination of $v_1, \ldots, v_n$, it follows that $\det B = 0$. But $B = sI - A$, where A is the matrix (a_{ij}). Thus s is a root of the monic polynomial $\det(xI - A) \in R[x]$ (the characteristic polynomial of A), and so s is a root of a monic polynomial with coefficients in R, which gives (1), completing the proof.

Corollary 24. Let $R \subseteq S$ be as in Proposition 23 and let $s, t \in S$.

(1) If s and t are integral over R then so are $s \pm t$ and st.

(2) The integral closure of R in S is a subring of S containing R.

(3) Integrality is transitive: let S be a subring of T; if T is integral over S and S is integral over R, then T is integral over R.

Proof: Let s and t be integral over R. By Proposition 23 both $R[s]$ and $R[t]$ are finitely generated R-modules, say

$$R[s] = Rs_1 + Rs_2 + \cdots + Rs_n$$
$$R[t] = Rt_1 + Rt_2 + \cdots + Rt_m.$$

Then

$$R[s,t] = Rs_1t_1 + \cdots + Rs_it_j + \cdots + Rs_nt_m$$

is a ring containing $s \pm t$ and st that is also a finitely generated R-module. Hence $s \pm t$ and st are also integral over R, which proves (1) and also (2).

To prove (3), let $t \in T$. Since t is integral over S, it is the root of some monic polynomial $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in S[x]$. Since $a_i \in S$ is integral over R, each ring $R[a_i]$ is a finitely generated R-module and so the ring $R_1 = R[a_0, a_1, \ldots, a_{n-1}]$ is also a finitely generated R-module. Since the monic polynomial $p(x)$ has its coefficients in R_1, t is integral over R_1 and it follows that the ring $R_1[t] = R[a_0, a_1, \ldots, a_{n-1}, t]$ is a finitely generated R-module. By the proposition, this means that t is integral over R, which gives (3).

The second statement in Corollary 24 shows that taking the elements of S that are integral over R gives a (possibly larger) subring of S, and the last statement in the corollary shows that the process of taking the integral closure stops after one step:

Corollary 25. Let R be a subring of the commutative ring S with $1 \in R$. Then the integral closure of R in S is integrally closed in S.

Examples

(1) If R and S are fields then S is integral over R if and only if S is algebraic over R — if $s \in S$ is a root of the polynomial $p(x)$ with coefficients in R then it is a root of the monic polynomial obtained by dividing by the (nonzero) leading coefficient of $p(x)$.

(2) Suppose S is an integral extension of R and I is an ideal in S. Then S/I is an integral ring extension of $R/(R \cap I)$ (reducing the monic polynomial over R satisfied by $s \in S$ modulo I gives a monic polynomial satisfied by $\bar{s} \in S/I$ over $R/(R \cap I)$).

(3) If R is a U.F.D. then R is integrally closed, as follows. Suppose a/b is an element in the field of fractions of R (with $b \neq 0$ and a and b having no common factors) and satisfies $(a/b)^n + r_{n-1}(a/b)^{n-1} + \cdots + r_1(a/b) + r_0 = 0$ with $r_0, \ldots, r_{n-1} \in R$. Then

$$a^n = b(-r_{n-1}a^{n-1} - \cdots - r_1ab^{n-2} - r_0b^{n-1})$$

shows that any irreducible element dividing b divides a^n, hence divides a. Since a/b is in lowest terms, this shows that b must be a unit, i.e., $a/b \in R$.

(4) The polynomial ring $k[x,y]$ over the field k is integrally closed in its fraction field $k(x,y)$ by example (3) above. The ideal $(x^2 - y^3)$ is prime (cf. Exercise 14, Section 9.1), so the quotient ring $R = k[x,y]/(x^2 - y^3) = k[\bar{x}, \bar{y}]$ is an integral domain. This domain is not integrally closed, however, since $\bar{x}/\bar{y}$ is an element of the fraction field of R that is integral over R (since $(\bar{x}/\bar{y})^3 - \bar{x} = 0$), but is not an element of R. In particular, R is not a U.F.D. by the previous example.

We next consider the behavior of ideals in integral ring extensions.

Definition. Let $\varphi : R \to S$ be a homomorphism of commutative rings.

(a) If I is an ideal in R then the *extension* of I to S is the ideal $\varphi(I)S$ of S generated by the image of I.

(b) If J is an ideal of S, then the *contraction* in R of J is the ideal $\varphi^{-1}(J)$.

In the special case where R is a subring of S and φ is the natural injection, the extension of $I \subseteq R$ is the ideal IS in S and the contraction of $J \subseteq S$ is the ideal $J \cap R$ of R.

It is immediate from the definition that

(1) $I \subseteq IS \cap R$, more generally, I is contained in the contraction of its extension to S, and

(2) $(J \cap R)\, S \subseteq J$, more generally, J contains the extension of its contraction in R.

In general equality need not hold in either situation (cf. the exercises).

If Q is a prime ideal in S, then its contraction is prime in R (although the contraction of a maximal ideal need not be maximal). On the other hand, if P is a prime ideal in R, its extension need not be prime (or even proper) in S; moreover, it is not generally true that P is the contraction of a prime ideal of S (cf. the exercises). For integral ring extensions, however, the situation is more controlled:

Theorem 26. Let R be a subring of the commutative ring S with $1 \in R$ and suppose that S integral over R.

(1) Assume that S is an integral domain. Then R is a field if and only if S is a field.

(2) Let P be a prime ideal in R. Then there is a prime ideal Q in S with $P = Q \cap R$. Moreover, P is maximal if and only if Q is maximal.

(3) *(The Going-up Theorem)* Let $P_1 \subseteq P_2 \subseteq \cdots \subseteq P_n$ be a chain of prime ideals in R and suppose there are prime ideals $Q_1 \subseteq Q_2 \subseteq \cdots \subseteq Q_m$ of S with $P_i = Q_i \cap R$, $1 \leq i \leq m$ and $m < n$. Then the ascending chain of ideals can be completed: there are prime ideals $Q_{m+1} \subseteq \cdots \subseteq Q_n$ in S such that $P_i = Q_i \cap R$ for all i.

(4) *(The Going-down Theorem)* Assume that S is an integral domain and R is integrally closed in S. Let $P_1 \supseteq P_2 \supseteq \cdots \supseteq P_n$ be a chain of prime ideals in R and suppose there are prime ideals $Q_1 \supseteq Q_2 \supseteq \cdots \supseteq Q_m$ of S with $P_i = Q_i \cap R$, $1 \leq i \leq m$ and $m < n$. Then the descending chain of ideals can be completed: there are prime ideals $Q_{m+1} \supseteq \cdots \supseteq Q_n$ in S such that $P_i = Q_i \cap R$ for all i.

Proof: To prove (1) assume first that R is a field and let s be a nonzero element of S. Then s is integral over R, so

$$s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 = 0$$

for some $a_0, a_1, \ldots, a_{n-1}$ in R. Since S is an integral domain, we may assume $a_0 \neq 0$ (otherwise cancel factors of s). Then

$$s(s^{n-1} + a_{n-1}s^{n-2} + \cdots + a_1) = -a_0$$

and since $(-1/a_0) \in R$, this shows that $(-1/a_0)(s^{n-1} + a_{n-1}s^{n-2} + \cdots + a_1)$ is an inverse for s in S, so S is a field. Conversely, suppose S is a field and r is a nonzero element of R. Since $r^{-1} \in S$ is integral over R we have

$$r^{-m} + a_{m-1}r^{-m+1} + \cdots + a_1 r^{-1} + a_0 = 0$$

for some $a_0, \dots, a_{m-1} \in R$. Then $r^{-1} = -(a_{m-1} + \cdots + a_1 r^{m-2} + a_0 r^{m-1}) \in R$, so R is a field.

The proof of the first statement in (2) is given in Corollary 50. For the second statement, observe that the integral domain S/Q is an integral extension of R/P (Example 2 following Corollary 25). By (1), S/Q is a field if and only if R/P is a field, i.e., Q is maximal if and only if P is maximal.

To prove (3), it suffices by induction to prove that if $P_1 \subseteq P_2$ and Q_1 is a prime of S with $Q_1 \cap R = P_1$ then there is a prime Q_2 of S with $Q_1 \subseteq Q_2$ and $Q_2 \cap R = P_2$. Since $\overline{S} = S/Q_1$ is an integral extension of $\overline{R} = R/P_1$, the first part of (2) shows that there exists a prime $\overline{Q_2}$ of $\overline{S}$ with $\overline{Q_2} \cap \overline{R} = P_2/P_1$. Then the preimage Q_2 of $\overline{Q_2}$ in S is a prime ideal containing Q_1 with $Q_2 \cap R = P_2$.

The proof of (4) is outlined in Exercise 24 in Section 4.

Corollary 27. Suppose R is a subring of the ring S with $1 \in R$ and assume S is integral and finitely generated (as a ring) over R. If P is a maximal ideal in R then there is a nonzero and finite number of maximal ideals Q of S with $Q \cap R = P$.

Proof: There exists at least one maximal ideal Q lying over P by (2) of the theorem, so we must see why there are only finitely many such maximal ideals in S. If Q is a maximal ideal of S with $Q \cap R = P$ then S/Q is a field containing the field R/P. To prove that there are only finitely many possible Q it suffices to prove that there are only finitely many homomorphisms from S to a field containing R/P that extend the homomorphism from R to R/P. Let $S = R[s_1, \dots, s_n]$, where the elements s_i are integral over R by assumption, and let $p_i(x)$ be a monic polynomial with coefficients in R satisfied by s_i. If Q is a maximal ideal of S then $S/Q = (R/P)[\bar{s}_1, \dots, \bar{s}_n]$ is the field extension of the field R/P with generators $\bar{s}_1, \dots, \bar{s}_n$. The element $\bar{s}_i$ is a root of the monic polynomial $\bar{p}_i(x)$ with coefficients in R/P obtained by reducing the coefficients of $p_i(x)$ mod P. There are only a finite number of possible roots of this monic polynomial (in a fixed algebraic closure of R/P), and so only finitely many possible field extensions of the form $(R/P)[\bar{s}_1, \dots, \bar{s}_n]$, which proves the corollary.

Algebraic Integers

We can use the concept of an integral ring extension to define the "integers" in extension fields of the rational numbers $\mathbb{Q}$:

Definition. Let K be an extension field of $\mathbb{Q}$.

(1) An element $\alpha \in K$ is called an *algebraic integer* if α is integral over $\mathbb{Z}$, i.e., if α is the root of some monic polynomial with coefficients in $\mathbb{Z}$.

(2) The integral closure of $\mathbb{Z}$ in K is called the *ring of integers* of K, and is denoted by $\mathcal{O}_K$.

An algebraic integer is clearly algebraic over $\mathbb{Q}$, so the ring of all algebraic integers is the ring of integers in $\overline{\mathbb{Q}}$, an algebraic closure of $\mathbb{Q}$. Examples of algebraic integers include $\sqrt{2}$, $\sqrt{-1}$, $\sqrt[3]{5}$, etc. since these elements are certainly roots of monic polynomials with coefficients in $\mathbb{Z}$. The definition of an algebraic integer α is that α be a root of *some* monic polynomial in $\mathbb{Z}[x]$, a condition which seems difficult to check. The next proposition gives a simple criterion for α to be an algebraic integer in terms of the minimal polynomial for α.

Proposition 28. An element α in some field extension of $\mathbb{Q}$ is an algebraic integer if and only if α is algebraic over $\mathbb{Q}$ and its minimal polynomial $m_{\alpha,\mathbb{Q}}(x)$ has integer coefficients. In particular, the algebraic integers in $\mathbb{Q}$ are the integers $\mathbb{Z}$, i.e., $\mathcal{O}_{\mathbb{Q}} = \mathbb{Z}$.

Proof: If α is algebraic over $\mathbb{Q}$ with $m_{\alpha,\mathbb{Q}}(x) \in \mathbb{Z}[x]$, then by definition α is integral over $\mathbb{Z}$. Conversely, assume α is integral over $\mathbb{Z}$, and let $f(x)$ be a monic polynomial in $\mathbb{Z}[x]$ of minimum degree having α as a root. If f were reducible in $\mathbb{Q}[x]$, then by Gauss' Lemma $f(x) = g(x)h(x)$ for some monic polynomials $g(x)$, $h(x)$ in $\mathbb{Z}[x]$ of degree smaller than the degree of f. But then α would be a root of either g or h, contradicting the minimality of f. Hence f is irreducible in $\mathbb{Q}[x]$, so $f(x) = m_{\alpha,\mathbb{Q}}(x)$ and so the minimal polynomial for α has coefficients in $\mathbb{Z}$. Finally, the minimal polynomial of $\alpha = a/b \in \mathbb{Q}$ (a/b reduced to lowest terms and $b > 0$) is $bx - a$, which is monic if and only if $b = 1$, so $\alpha \in \mathbb{Q}$ is an algebraic integer if and only if $\alpha \in \mathbb{Z}$.

Because the integers $\mathbb{Z}$ are the algebraic integers in $\mathbb{Q}$, for emphasis (and clarity) the elements of $\mathbb{Z}$ are sometimes referred to as the "rational integers" to distinguish them from the "integers" in extensions of finite degree over $\mathbb{Q}$ (called *number fields*). The next result gives some of the basic structure of the ring of integers in a general number field.

Theorem 29. Let K be a number field of degree n over $\mathbb{Q}$.

(1) The ring $\mathcal{O}_K$ of integers in K is a Noetherian ring and is a free $\mathbb{Z}$-module of rank n.

(2) For every $\beta \in K$ there is some nonzero $d \in \mathbb{Z}$ such that $d\beta$ is an algebraic integer. In particular, K is the field of fractions of $\mathcal{O}_K$.

(3) If $\beta_1, \beta_2, \dots, \beta_n$ is any $\mathbb{Q}$-basis of K, then there is an integer d such that $d\beta_1, d\beta_2, \dots, d\beta_n$ is a basis for a free $\mathbb{Z}$-submodule of $\mathcal{O}_K$ of rank n. Any basis of the $\mathbb{Z}$-module $\mathcal{O}_K$ is also a basis for K as a vector space over $\mathbb{Q}$.

Proof: Note first that any $\mathbb{Z}$-linear dependence relation among elements in $\mathcal{O}_K$ is a $\mathbb{Q}$-linear dependence relation in K, and multiplying a $\mathbb{Q}$-linear dependence relation of elements of $\mathcal{O}_K$ in K by a common denominator for the coefficients yields a $\mathbb{Z}$-linear dependence relation in $\mathcal{O}_K$. Let β be any element of K and let $x^k + a_{k-1}x^{k-1} + \cdots + a_0$ be the minimal polynomial of β over $\mathbb{Q}$. If d is a common denominator for the coefficients, then multiplying through by d^k shows that

$$(d\beta)^k + da_{k-1}(d\beta)^{k-1} + \cdots + d^{k-1}a_1(d\beta) + d^k a_0 = 0,$$

and $d^k a_0, d^{k-1}a_1, \dots, da_{k-1} \in \mathbb{Z}$. Hence $d\beta$ is an algebraic integer, which proves the first part of (2) and then the second statement in (2) follows immediately.

If $\beta_1, \dots, \beta_n$ are a $\mathbb{Q}$-basis for K over $\mathbb{Q}$, then there is a nonzero integer d such that $d\beta_1, \dots, d\beta_n$ all lie in $\mathcal{O}_K$. These elements are still linearly independent over $\mathbb{Q}$, so in particular are independent over $\mathbb{Z}$, hence generate a free submodule of $\mathcal{O}_K$ of rank n, which proves the first statement in (3).

Since $\mathcal{O}_K$ is a subring of the field K, it is a torsion free $\mathbb{Z}$-module. If $\mathcal{O}_K$ were contained in some finitely generated $\mathbb{Z}$-module it would follow that $\mathcal{O}_K$ is also finitely generated over $\mathbb{Z}$, hence is a free $\mathbb{Z}$-module. If L is the Galois closure of K, then $\mathcal{O}_K \subseteq \mathcal{O}_L$ and so it suffices to see that $\mathcal{O}_L$ is contained in a finitely generated $\mathbb{Z}$-module. Let $\alpha_1, \dots, \alpha_m$ be a $\mathbb{Q}$-basis for L. Multiplying by an integer $d \in \mathbb{Z}$, if necessary, we may assume that each α_i is an algebraic integer, i.e., $\alpha_1, \dots, \alpha_m \in \mathcal{O}_L$. For each fixed $\theta \neq 0$ in L, the map

$$T_\theta : L \to \mathbb{Q} \quad \text{defined by} \quad T_\theta(\alpha) = \operatorname{Tr}_{L/\mathbb{Q}}(\theta\alpha)$$

(where $\operatorname{Tr}_{L/\mathbb{Q}}$ denotes the trace map from L to $\mathbb{Q}$, cf. Exercise 18 in Section 14.2) is a $\mathbb{Q}$-linear transformation from L to $\mathbb{Q}$. This linear transformation is nonzero because $T_\theta(\theta^{-1}) = \operatorname{Tr}_{L/\mathbb{Q}}(1) = m$. It follows that the map from L to $\operatorname{Hom}_{\mathbb{Q}}(L, \mathbb{Q})$ mapping θ to T_θ is an injective homomorphism of vector spaces over $\mathbb{Q}$. Since both spaces have the same dimension over $\mathbb{Q}$, the map is an isomorphism. Put another way, every linear functional on L is of the form T_θ for some $\theta \in L$. In particular, there are elements $\alpha'_1, \dots, \alpha'_m$ in L whose corresponding linear transformations $T_{\alpha'_i}$ give the dual basis of $\alpha_1, \dots, \alpha_m$, i.e.,

$$\operatorname{Tr}_{L/\mathbb{Q}}(\alpha'_i\alpha_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

Since $\alpha'_1, \dots, \alpha'_m$ are linearly independent, they give a basis for L over $\mathbb{Q}$. Hence every element $\beta \in \mathcal{O}_L$ can be written

$$\beta = a_1\alpha'_1 + \cdots + a_i\alpha'_i + \cdots + a_m\alpha'_m$$

with $a_1, \dots, a_m \in \mathbb{Q}$. Multiplying by α_j and taking the trace shows that

$$\operatorname{Tr}_{L/\mathbb{Q}}(\beta\alpha_j) = a_1\operatorname{Tr}_{L/\mathbb{Q}}(\alpha'_1\alpha_j) + \cdots + a_i\operatorname{Tr}_{L/\mathbb{Q}}(\alpha'_i\alpha_j) + \cdots + a_m\operatorname{Tr}_{L/\mathbb{Q}}(\alpha'_m\alpha_j) = a_j.$$

But β and α_j are both elements of $\mathcal{O}_L$, so also $\beta\alpha_j$ is an element of $\mathcal{O}_L$, and this implies that $a_j = \operatorname{Tr}_{L/\mathbb{Q}}(\beta\alpha_j)$ is an element of $\mathbb{Z}$ (cf. Exercise 18(d) of Section 14.2). It follows that

$$\mathcal{O}_L \subseteq \mathbb{Z}\alpha'_1 + \cdots + \mathbb{Z}\alpha'_m$$

so that $\mathcal{O}_L$ is contained in a finitely generated $\mathbb{Z}$-module, proving that $\mathcal{O}_K$ (and also $\mathcal{O}_L$) is a free $\mathbb{Z}$-module.

Since K has dimension n as a vector space over $\mathbb{Q}$, it follows that $\mathcal{O}_K$ is a free $\mathbb{Z}$-module of rank at most n (by Theorem 5 of Section 12.1). Because $\mathcal{O}_K$ also contains a free $\mathbb{Z}$-submodule of rank n, it follows that the $\mathbb{Z}$-rank of $\mathcal{O}_K$ is precisely n, proving (1), and then the second statement in (3) follows by the remarks on $\mathbb{Z}$-linear and $\mathbb{Q}$-linear dependence relations.

Finally, any ideal I in $\mathcal{O}_K$ is a $\mathbb{Z}$-submodule of a free $\mathbb{Z}$-module of rank n, so is a free $\mathbb{Z}$-module of rank at most n, and a set of $\mathbb{Z}$-module generators for I is also a set of $\mathcal{O}_K$-generators. Hence every ideal of $\mathcal{O}_K$ can be generated by at most n elements, which implies that $\mathcal{O}_K$ is a Noetherian ring and completes the proof.

Definition. An *integral basis* for the number field K is a basis of the ring of integers in K considered as a free $\mathbb{Z}$-module of rank $[K : \mathbb{Q}]$.

If P is a nonzero prime ideal in the ring of integers $\mathcal{O}_K$ of a number field K then $P \cap \mathbb{Z}$ is a prime ideal in $\mathbb{Z}$. If $\alpha \in P$, then the constant term of the minimal polynomial for α over $\mathbb{Q}$ is then an element in $P \cap \mathbb{Z}$, which shows that $P \cap \mathbb{Z} = p\mathbb{Z}$ is also a nonzero prime ideal in $\mathbb{Z}$. By Theorem 26, every prime ideal (p) in $\mathbb{Z}$ arises in this way. Since $p\mathbb{Z}$ is a maximal ideal, it also follows from (2) in Theorem 26 that *nonzero prime ideals in $\mathcal{O}_K$ are maximal*, and then by Corollary 27, there are finitely many prime ideals P in $\mathcal{O}_K$ with $P \cap \mathbb{Z} = p\mathbb{Z}$. We shall see later (Corollary 16 in Section 16.3) that *every nonzero ideal in the ring of integers of a number field can be written uniquely as the product of prime ideals*, and in the case of the ideal $p\mathcal{O}_K$ the distinct prime factors are precisely the finitely many ideals P in $\mathcal{O}_K$ with $P \cap \mathbb{Z} = p\mathbb{Z}$. This property replaces the unique factorization of *elements* in $\mathcal{O}_K$ into primes (which need not hold since $\mathcal{O}_K$ *need not be a U.F.D.*). We shall also see that primary ideals in $\mathcal{O}_K$ are powers of prime ideals (in fact this is equivalent to the unique factorization of ideals of $\mathcal{O}_K$ into products of prime ideals, cf. the exercises).

Example: (The Ring of Integers in Quadratic Extensions of $\mathbb{Q}$)

If K is a quadratic extension of $\mathbb{Q}$ then $K = \mathbb{Q}(\sqrt{D})$ for some squarefree integer D. Then

$$\mathcal{O}_{\mathbb{Q}(\sqrt{D})} = \mathbb{Z}[\omega] = \mathbb{Z} \cdot 1 + \mathbb{Z} \cdot \omega,$$

with integral basis $1, \omega$, where

$$\omega = \begin{cases} \sqrt{D}, & \text{if } D \equiv 2, 3 \bmod 4 \\ \dfrac{1+\sqrt{D}}{2}, & \text{if } D \equiv 1 \bmod 4. \end{cases}$$

This is the quadratic integer ring introduced in Section 7.1. Since ω satisfies $\omega^2 - D = 0$ (respectively, $\omega^2 - \omega + (1 - D)/4$) for $D \equiv 2, 3 \bmod 4$ (respectively, $D \equiv 1 \bmod 4$), it follows that ω is an algebraic integer in K and so $\mathbb{Z}[\omega] \subseteq \mathcal{O}_K$. To prove that this is the full ring of integers in K, let $\alpha = a + b\sqrt{D}$ with $a, b \in \mathbb{Q}$, and suppose that α is an algebraic integer. If $b = 0$, then $\alpha \in \mathbb{Q}$ and so $a \in \mathbb{Z}$. If $b \neq 0$, the minimal polynomial of α is $x^2 - 2ax + (a^2 - b^2D)$. Then Proposition 28 shows that $2a$ and $a^2 - b^2D$ are elements of $\mathbb{Z}$. Then $4(a^2 - b^2D) = (2a)^2 - (2b)^2D \in \mathbb{Z}$, hence $4b^2D \in \mathbb{Z}$. Since D is squarefree it follows that $2b$ is an integer. Write $a = x/2$ and $b = y/2$ for some integers x, y. Since $a^2 - b^2D$ is an integer, $x^2 - y^2D \equiv 0 \pmod 4$. Since 0 and 1 are the only squares mod 4 and D is not divisible by 4, it is easy to check that the only possibilities are the following:

(i) $D \equiv 2$ or $3 \pmod 4$ and x, y are both even, or

(ii) $D \equiv 1 \pmod 4$ and x, y are both even or both odd.

In case (i), $a, b \in \mathbb{Z}$ and $\alpha \in \mathbb{Z}[\omega]$. In case (ii), $a + b\sqrt{D} = r + s\omega$ where $r = (x - y)/2$ and $s = y$ are both integers, so again $\alpha \in \mathbb{Z}[\omega]$.

Example: (The Ring of Integers in Cyclotomic Fields)

The ring of integers in the cyclotomic field $\mathbb{Q}(\zeta_n)$ of n^{th} roots of unity is $\mathbb{Z}[\zeta_n]$, where ζ_n is any primitive n^{th} root of 1. The elements $1, \zeta_n, \ldots, \zeta_n^{\varphi(n)-1}$ are an integral basis. It is clear that ζ_n is an algebraic integer since it is a root of $x^n - 1$, so the ring $\mathbb{Z}[\zeta_n]$ is contained in the ring of integers. The proof that this is the full ring of algebraic integers in $\mathbb{Q}(\zeta_n)$ involves techniques from algebraic number theory beyond the scope of the material here.

Noether's Normalization Lemma and Hilbert's Nullstellensatz

We now apply some of the techniques from the algebraic theory of integral ring extensions to affine geometry.

Definition. If k is a field the elements $y_1, y_2, \ldots, y_q$ in some k-algebra are called *algebraically independent* over k if there is no nonzero polynomial p in q variables over k such that $p(y_1, y_2, \ldots, y_q) = 0$.

Thus $y_1, y_2, \ldots, y_q$ are algebraically independent if and only if the k-algebra homomorphism from the polynomial ring $k[x_1, \ldots, x_q]$ to $k[y_1, \ldots, y_q]$ defined by $x_i \mapsto y_i$ is an isomorphism. Elements in a field extension of k are algebraically independent if and only if they are independent transcendentals over k.

Theorem 30. *(Noether's Normalization Lemma)* Let k be a field and suppose that $A = k[r_1, r_2, \ldots, r_m]$ is a finitely generated k-algebra. Then for some q, $0 \leq q \leq m$, there are algebraically independent elements $y_1, y_2, \ldots, y_q \in A$ such that A is integral over $k[y_1, y_2, \ldots, y_q]$.

Proof: Proceed by induction on m. If $r_1, \ldots, r_m$ are algebraically independent over k then take $y_i = r_i$, $i = 1, \ldots, m$. Otherwise, there exists $f(x_1, \ldots, x_m) \in k[x_1, \ldots, x_m]$ such that $f(r_1, \ldots, r_m) = 0$. The polynomial f is a sum of monomials of the form $ax_1^{e_1}x_2^{e_2} \cdots x_m^{e_m}$, where the degree of this monomial is $e_1 + \cdots + e_m$ and the degree, d, of f is the maximum of the degrees of its monomials. Renumbering the variables if necessary, we may assume that f is a nonconstant polynomial in x_m with coefficients in the ring $k[x_1, x_2, \ldots, x_{m-1}]$. We now perform a change of variables that transforms (or "normalizes") f into a *monic* polynomial in x_m with coefficients from a subring of A which is generated over k by $m - 1$ elements, at which point we shall be able to apply induction.

Define integers $\alpha_i = (1 + d)^i$ and new variables $X_i = x_i - x_m^{\alpha_i}$ for $1 \leq i \leq m - 1$. Let

$$g(X_1, X_2, \ldots, X_{m-1}, x_m) = f(X_1 + x_m^{\alpha_1}, X_2 + x_m^{\alpha_2}, \ldots, X_{m-1} + x_m^{\alpha_{m-1}}, x_m),$$

so $g \in k[X_1, \ldots, X_{m-1}, x_m]$. Each monomial term of f contributes a single term of the form a constant times x_m^e to g. It is also easy to check that the choice of α_i ensures that distinct monomials in f give different values of e (for example by viewing the degrees of the monomials in the new variables as integers expressed in base $b = d + 1$). If N is the highest power of x_m that occurs, then it follows that

$$g = cx_m^N + \sum_{i=0}^{N-1} h_i(X_1, \ldots, X_{m-1})x_m^i$$

for some nonzero $c \in k$. If now $s_i = r_i - r_m^{\alpha_i}$ then

$$\frac{1}{c}g(s_1, s_2, \ldots, s_{m-1}, r_m) = \frac{1}{c}f(r_1, r_2, \ldots, r_{m-1}, r_m) = 0,$$

which shows that r_m is integral over $B = k[s_1, \ldots, s_{m-1}]$. Each r_i for $1 \leq i \leq m - 1$ is integral over $B[r_m]$ since r_i is a root of the monic polynomial $x - s_i - r_m^{\alpha_i}$, so A is integral over $B[r_m]$. By transitivity of integrality, A is therefore integral over B. Since B is a k-algebra generated by $m-1$ elements, induction completes the proof.

A more "geometric" interpretation of Noether's Normalization Lemma is indicated in Exercise 15. We next use the Normalization Lemma to prove that if k is an algebraically closed field then the maximal ideals of the polynomial ring $k[x_1, x_2, \dots, x_n]$ are of the form $(x_1-a_1, \dots, x_n-a_n)$ for some $a_1, \dots, a_n \in k$. Viewing $k[x_1, x_2, \dots, x_n]$ as the ring of polynomial functions on $\mathbb{A}^n$, this says that the maximal ideals correspond to the kernels of evaluation maps at points of $\mathbb{A}^n$ — similar to the corresponding result for rings of continuous functions on a compact set (cf. Exercises 33, 34 in Section 7.4).

Theorem 31. *(Hilbert's Nullstellensatz — Weak Form)* Let k be an algebraically closed field. Then M is a maximal ideal in the polynomial ring $k[x_1, x_2, \dots, x_n]$ if and only if $M = (x_1 - a_1, \dots, x_n - a_n)$ for some $a_1, \dots, a_n \in k$. Equivalently, the maps $\mathcal{Z}$ and $\mathcal{I}$ give a bijective correspondence

$$\{\text{points in } \mathbb{A}^n\} \quad \underset{\mathcal{Z}}{\overset{\mathcal{I}}{\rightleftarrows}} \quad \{\text{maximal ideals in } k[\mathbb{A}^n]\}.$$

Moreover, if I is any proper ideal in $k[x_1, x_2, \dots, x_n]$ then $\mathcal{Z}(I) \neq \emptyset$.

Proof: Certainly $(x_1 - a_1, \dots, x_n - a_n)$ is a maximal ideal in $k[x_1, x_2, \dots, x_n]$. Conversely, for any maximal ideal M in $k[x_1, x_2, \dots, x_n]$, let $E = k[x_1, x_2, \dots, x_n]/M$. Then E is a field containing k that is finitely generated over k (by $\bar{x}_i, \dots, \bar{x}_n$). By Noether's Normalization Lemma, E is integral over a polynomial ring $k[y_1, \dots, y_q]$. Then $k[y_1, \dots, y_q]$ is a field by Theorem 26(1), and since a polynomial ring in one or more variables is never a field, it follows that $q = 0$. Hence E is integral over k, so E is algebraic over k. Because k is algebraically closed, $E = k$, i.e., $\bar{x}_i \in k$ for $1 \leq i \leq n$. Hence for $i = 1, \dots, n$ there is some $a_i \in k$ such that $x_i - a_i \in M$. This means that the maximal ideal $(x_1 - a_1, \dots, x_n - a_n)$ is contained in M, so $M = (x_1 - a_1, \dots, x_n - a_n)$. Finally, if I is any nonzero ideal in $k[x_1, x_2, \dots, x_n]$ then I is contained in a maximal ideal $M = (x_1 - a_1, \dots, x_n - a_n)$, and so $(a_1, \dots, a_n) \in \mathcal{Z}(I)$.

Theorem 32. *(Hilbert's Nullstellensatz)* Let k be an algebraically closed field. Then $\mathcal{I}(\mathcal{Z}(I)) = \operatorname{rad} I$ for every ideal I of $k[x_1, x_2, \dots, x_n]$. Moreover, the maps $\mathcal{Z}$ and $\mathcal{I}$ define inverse bijections

$$\{\text{affine algebraic sets}\} \quad \underset{\mathcal{Z}}{\overset{\mathcal{I}}{\rightleftarrows}} \quad \{\text{radical ideals}\}.$$

Proof: Since $\operatorname{rad} I \subseteq \mathcal{I}(\mathcal{Z}(I))$ it remains to prove the reverse inclusion. By Hilbert's Basis Theorem, $I = (f_1, f_2, \dots, f_m)$. Let $g \in \mathcal{I}(\mathcal{Z}(I))$. Introduce a new variable x_{n+1} and consider the ideal I' generated by $f_1, \dots, f_m$ and $x_{n+1}g - 1$ in $k[x_1, \dots, x_n, x_{n+1}]$. At any point of $\mathbb{A}^{n+1}$ where $f_1, \dots, f_m$ vanish the polynomial g also vanishes since $g \in \mathcal{I}(\mathcal{Z}(I))$, so that $x_{n+1}g - 1$ is nonzero. Hence $\mathcal{Z}(I') = \emptyset$ in $\mathbb{A}^{n+1}$. By the Weak Form of the Nullstellensatz, I' cannot be a proper ideal, i.e., $1 \in I'$. Write

$$1 = a_1 f_1 + \cdots + a_m f_m + a_{m+1}(x_{n+1}g - 1) \qquad \text{for some } a_i \in k[x_1, \dots, x_{n+1}].$$

Letting $y = 1/x_{n+1}$ and multiplying by a high power of y in this equation shows that

$$y^N = c_1 f_1 + \cdots + c_m f_m + c_{m+1}(g - y) \qquad \text{for some } c_i \in k[x_1, \dots, x_n, y].$$

Substituting g for y in this polynomial equation shows that $g^N \in I$ (in $k[x_1, \dots, x_n]$), i.e., $g \in \operatorname{rad} I$. Hence $\mathcal{I}(\mathcal{Z}(I)) \subseteq \operatorname{rad} I$ and so $\mathcal{I}(\mathcal{Z}(I)) = \operatorname{rad} I$, completing the proof.

It follows directly from Proposition 12 and Theorem 26(2) that if S is an integral extension of R with $1 \in R$ and if I is an ideal of R, then

$$(\operatorname{rad}_S IS) \cap R = \operatorname{rad}_R I$$

where IS is the ideal generated by I in S, and the subscript indicates the ring in which the radicals are being computed. This has the following geometric interpretation.

Corollary 33. *(Variant of Hilbert's Nullstellensatz)* If k is any field with algebraic closure $\bar{k}$ and I is an ideal in $k[x_1, x_2, \dots, x_n]$, then $\mathcal{I}_k(\mathcal{Z}_{\bar{k}}(I)) = \operatorname{rad} I$, where $\mathcal{Z}_{\bar{k}}(I)$ is the zero set in $\bar{k}^n$ of the polynomials in I and $\mathcal{I}_k(\mathcal{Z}_{\bar{k}}(I))$ is the ideal of polynomials in $k[x_1, x_2, \dots, x_n]$ vanishing at all the points in $\mathcal{Z}_{\bar{k}}(I)$. In particular, $I = (1)$ if and only if there are no common zeros in $\bar{k}^n$ of the polynomials in I.

Proof: Since $\bar{k}[x_1, x_2, \dots, x_n]$ is an integral extension of $k[x_1, x_2, \dots, x_n]$ (generated by the integral elements $\bar{k}$), the corollary follows immediately from Theorem 32 and the remarks on radicals above.

From the Nullstellensatz we now have a dictionary between geometric and ring-theoretic objects over the algebraically closed field k:

Geometry	Algebra
affine algebraic set V	coordinate ring $k[V]$
points of V	maximal ideals of $k[V]$
affine algebraic subsets in V	radical ideals of $k[V]$
subvarieties in V	prime ideals in $k[V]$
morphism $\varphi : V \to W$	k-algebra homomorphism $\widetilde{\varphi} : k[W] \to k[V]$

Computing Radicals

There are algorithms for computing radicals and primary decompositions in polynomial rings using Gröbner bases. While they are relatively elementary, they are somewhat technical and so we limit our discussion here to some preliminary results.

For hypersurfaces $V = \mathcal{Z}(f)$ defined by a single polynomial $f \in k[x_1, \dots, x_n]$, determining $\mathcal{I}(V) = \operatorname{rad}(f)$ is straightforward. Since $k[x_1, \dots, x_n]$ is a U.F.D., f factors uniquely as the product of powers of nonassociate irreducibles: $f = p_1^{a_1} \cdots p_s^{a_s}$ and then $\operatorname{rad}(f)$ is generated by $p_1 \cdots p_s$ (the 'squarefree part' of f).

Example

Suppose $W = \mathcal{Z}(J)$ with $J = (u^3 - uv^2 + v^3) \in \mathbb{Q}[u, v]$. The polynomial $x^3 - x + 1$ is irreducible over $\mathbb{Q}$, so $f = u^3 - uv^2 + v^3$ is irreducible in $\mathbb{Q}[u, v]$. Hence rad $J = J$ and $\mathcal{I}(W) = J$.

For nonprincipal ideals I, determining rad I is more complicated. The following proposition (based on Hilbert's Nullstellensatz) gives a criterion determining when an element is contained in rad I.

Proposition 34. Suppose k is any field. If $I = (f_1, \dots, f_s)$ is a proper ideal in $k[x_1, \dots, x_n]$, then $f \in \text{rad}\, I$ if and only if $(f_1, \dots, f_s, 1 - yf) = k[x_1, \dots, x_n, y]$.

Proof: By Corollary 33, $(f_1, \dots, f_s, 1 - yf) = k[x_1, \dots, x_n, y]$ if and only if the equations

$$1 - yf(x_1, \dots, x_n) = 0, \quad f_1(x_1, \dots, x_n) = 0, \quad \dots, \quad f_s(x_1, \dots, x_n) = 0$$

have no common zero over the algebraic closure $\bar{k}$ of k. For a given $(a_1, \dots, a_n) \in \bar{k}^n$, the equation $1 - yf(a_1, \dots, a_n) = 0$ has a solution y unless $f(a_1, \dots, a_n) = 0$. Hence, the system of equations has no common zero if and only if for every $(a_1, \dots, a_n) \in \bar{k}^n$ with $f_1(a_1, \dots, a_n) = \cdots = f_s(a_1, \dots . a_n) = 0$ we also have $f(a_1, \dots, a_n) = 0$. Equivalently, if $(a_1, \dots, a_n) \in \mathcal{Z}_{\bar{k}}(I)$, then also $f(a_1, \dots, a_n) = 0$, i.e., we have $f \in \mathcal{I}_k(\mathcal{Z}_{\bar{k}}(I)) = \text{rad}\, I$, by Corollary 33.

Since the reduced Gröbner basis (with respect to any fixed monomial ordering) for an ideal is unique, we immediately obtain the following algorithmic method for determining when a polynomial lies in the radical of an ideal.

Corollary 35. Suppose $I = (f_1, \dots, f_s)$ in $k[x_1, \dots, x_n]$. Then $f \in \text{rad}\, I$ if and only if $\{1\}$ is the reduced Gröbner basis for the ideal $(f_1, \dots, f_s, 1 - yf)$ in $k[x_1, \dots, x_n, y]$ with respect to any monomial ordering.

Example

Consider $I = (x^2 - y^2, xy)$ in $k[x, y]$. The reduced Gröbner basis for $(x^2 - y^2, xy, 1 - tx)$ in $k[x, y, t]$ with respect to the order $x > y > t$ is $\{1\}$, showing $x \in \text{rad}(I)$. To determine the smallest power of x lying in I, we find that the ideal $(x^2 - y^2, xy, x^3)$ in $k[x, y]$ has the same reduced Gröbner basis as I (namely $\{x^2 - y^2, xy, y^3\}$), but $(x^2 - y^2, x^2, xy)$ has basis $\{x^2, xy, y^2\}$. It follows that $x^3 \in I$ and $x^2 \notin I$ (alternatively, x^3 leaves a nonzero remainder after general polynomial division by $\{x^2 - y^2, xy, y^3\}$, but x^3 has a remainder of 0). By a similar computation (or by symmetry), $y \in \text{rad}\, I$, with $y^3 \in I$ but $y^2 \notin I$. Since $(x, y) \subseteq \text{rad}\, I$, it follows that rad $I = (x, y)$.

Some additional results for computing radicals are presented in the exercises.

EXERCISES

Let R be a subring of the commutative ring S with $1 \in R$.

1. Use the fact that a U.F.D. is integrally closed to prove that the Gaussian integers, $\mathbb{Z}[i]$, is the ring of integers in $\mathbb{Q}(i)$.

2. Suppose k is a field and let $t = \bar{x}/\bar{y}$ in the field of fractions of the integral domain $R = k[x, y]/(x^2 - y^3)$. Prove that $K = k(t)$ is the fraction field of R and $k[t]$ is the integral closure of R in K.

3. Suppose k is a field and i and j are relatively prime positive integers. Find the normalization of the integral domain $R = k[x, y]/(x^i - y^j)$ (cf. Exercise 14, Section 9.1).

4. Suppose k is a field and let P be the ideal $(y^2 - x^3 - x^2)$ in the polynomial ring $k[x, y]$. Prove that P is a prime ideal and find the normalization of the integral domain $R = k[x, y]/P$. [To prove P is prime, show that $y^2 - x^3 - x^2$ is irreducible in the U.F.D. $k[x, y]$. Then consider $t = \bar{y}/\bar{x} \in R$.]

5. If R is an integral domain with field of fractions F, show that F is a finitely generated R-module if and only if $R = F$.

6. For each of the following give specific rings $R \subseteq S$ and explicit ideals in these rings that exhibit the specified relation:
   (a) an ideal I of R such that $I \neq SI \cap R$ (so the contraction of the extension of an ideal I need not equal I)
   (b) a prime ideal P of R such that there is no prime ideal Q of S with $P = Q \cap R$
   (c) a maximal ideal M of S such that $M \cap R$ is not maximal in R
   (d) a prime ideal P of R whose extension PS to S is not a prime ideal in S
   (e) an ideal J of S such that $J \neq (J \cap R)S$ (so the extension of the contraction of an ideal J need not equal J).

7. Let $\mathcal{O}_K$ be the ring of integers in a number field K.
   (a) Suppose that every nonzero ideal I of $\mathcal{O}_K$ can be written as the product of powers of prime ideals. Prove that an ideal Q of $\mathcal{O}_K$ is P-primary if and only if $Q = P^m$ for some $m \geq 1$. [Show first that since nonzero primes in $\mathcal{O}_K$ are maximal that ${P_1}^{m_1} \subseteq {P_2}^{m_2}$ for distinct nonzero primes P_1, P_2 implies $P_1 = P_2$.]
   (b) Suppose that an ideal Q of $\mathcal{O}_K$ is P-primary if and only if $Q = P^m$ for some $m \geq 1$. Assuming all of Theorem 21, prove that every nonzero ideal I of $\mathcal{O}_K$ can be written uniquely as the product of powers of prime ideals. [Prove that ${P_1}^{m_1}$ and ${P_2}^{m_2}$ are comaximal ideals if P_1 and P_2 are distinct nonzero prime ideals and use the Chinese Remainder Theorem.]

8. Prove that if $s_1, \dots, s_n \in S$ are integral over R, then the ring $R[s_1, \dots, s_n]$ is a finitely generated R-module.

9. Suppose that S is integral over R and that P is a prime ideal in R. Prove that every element s in the ideal PS generated by P in S satisfies an equation $s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 = 0$ where the coefficients $a_0, a_1, \dots, a_{n-1}$ are elements of P. [If $s = p_1 s_1 + \cdots + p_m s_m \in PS$, show that $T = R[s_1, \dots, s_m]$ satisfies the hypotheses in Proposition 23(3). Follow the proof in Proposition 23 that s is integral, noting that $s \in PT$ so that the a_{ij} are elements of P.]

10. Prove the following generalization of Proposition 28: Suppose R is an integrally closed integral domain with field of fractions k and α is an element of an extension field K of k. Show that α is integral over R if and only if α is algebraic over k and the minimal polynomial $m_{\alpha,k}(x)$ for α over k has coefficients in R. [If α is integral prove the conjugates

of α, i.e., the roots of $m_{\alpha,k}(x)$, are also integral, so the elementary symmetric functions of the conjugates are elements of k that are integral over R.]

11. Suppose R is an integrally closed integral domain with field of fractions k and $p(x) \in R[x]$ is a monic polynomial. Show that if $p(x) = a(x)b(x)$ with monic polynomials $a(x), b(x) \in k[x]$ then $a(x), b(x) \in R[x]$ (compare to Gauss' Lemma, Proposition 5, Section 9.3). [See the previous exercise.]

12. Suppose S is an integral domain that is integral over a ring R as in the previous exercise. If P is a prime ideal in R, let s be any element in the ideal PS generated by P in S. Prove that, with the exception of the leading term, the coefficients of the minimal polynomial $m_{s,k}(x)$ for s over k are elements of P. [By Exercise 10, $m_{s,k}(x) \in R[x]$. Exercise 9 shows that s is a root of a monic polynomial $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ with $a_0, \dots, a_{n-1} \in P$. Use the previous exercise to show that $p(x) = m_{s,k}(x)b(x)$ with $b(x)$ in $R[x]$, and consider this equation in the integral domain $(R/P)[x]$.]

The next two exercises extend Exercise 6 in Section 7.5 by characterizing fields that are not fields of fractions of any of their proper subrings.

13. Let K be a field of characteristic 0 and let A be a subring of K maximal with respect to $1/2 \notin A$. (Such A exists by Zorn's Lemma.) Let F be the field of fractions of A in K.
   (a) Show that K is algebraic over F. [If t is transcendental over F, show that $1/2 \notin A[t]$.]
   (b) Show that A is integrally closed in K. [Show that $1/2$ is not in the integral closure of A in K.]
   (c) Deduce from (a) and (b) that $K = F$.

14. Show that a field K is the field of fractions of some proper subring of K if and only if K is not a subfield of the algebraic closure of a finite field. [If K contains t transcendental over $\mathbb{F}_p$ argue as in the preceding exercise with $1/t$ in place of $1/2$ to show that K is the quotient field of some proper subring.]

The next exercise gives a "geometric" interpretation of Noether's Normalization Lemma, showing that every affine algebraic set is a *finite covering* of some affine n-space.

15. Let V be an affine algebraic set over an algebraically closed field k. Prove that for some n there is a surjective morphism from V onto $\mathbb{A}^n$ with finite fibers, and that if V is a variety, then n can be taken to be the dimension of V. [By Noether's Normalization Lemma the finitely generated k-algebra $S = k[V]$ contains a polynomial subalgebra $R = k[x_1, x_2, \dots, x_n]$ such that S is integral over R. Apply Theorem 6 to the inclusion of R in S to obtain a morphism φ from V to $\mathbb{A}^n$. To see that φ is surjective with finite fibers, apply Corollary 27 to the maximal ideal $(x_1 - a_1, \dots, x_n - a_n)$ of R corresponding to a point $(a_1, \dots, a_n)$ of $\mathbb{A}^n$.]

16. Let V be an affine algebraic set in $\mathbb{C}^n$. Prove that V is compact in the Euclidean topology (i.e., closed and bounded) if and only if it is finite. [Use Exercise 18 in Section 2, the previous exercise, and the behavior of compact sets with respect to continuous functions.]

17. Let R be a subring of the commutative ring S with $1_S \in R$ and suppose that S is integral over R. This exercise proves that R and S have the same *Krull dimension*, cf. Section 16.1.
   (a) If $P_1 \subset P_2 \subset \cdots \subset P_n$ is a chain of distinct prime ideals in R prove that there is a chain $Q_1 \subset Q_2 \subset \cdots \subset Q_n$ of distinct prime ideals in S with $Q_i \cap R = P_i$.
   (b) Prove conversely that if $Q_1 \subset Q_2 \subset \cdots \subset Q_n$ is a chain of distinct prime ideals in S and $P_i = Q_i \cap R$ then $P_1 \subset P_2 \subset \cdots \subset P_n$ is a chain of distinct prime ideals in R. [To prove the P_i are distinct, pass to a quotient and reduce the problem to showing that if Q is a nonzero prime ideal in the integral domain S then $Q \cap R$ is a nonzero prime

ideal in R. In this case, if $s \in Q$ is nonzero, show that the constant coefficient of a polynomial of minimal degree in $R[x]$ satisfied by s is a nonzero element in $Q \cap R$.]

18. Let $V = \mathcal{Z}(I)$ and $W = \mathcal{Z}(J)$ where I is the ideal $(uv + v) \subset \mathbb{C}[u, v]$ and J is the ideal $(-2y - y^2 + 2z + z^2, 2x - yz - z^2) \subset \mathbb{C}[x, y, z]$.

(a) Show that I and J are prime ideals. Conclude that $I = \mathcal{I}(V)$ and $J = \mathcal{I}(W)$ and that V and W are varieties.

(b) Show that the map $\varphi : V \to W$ defined by $\varphi((a_1, a_2)) = (a_1^2 + a_2, a_1 + a_2, a_1 - a_2)$ is an isomorphism.

19. Let $I = (x^3 + y^3 + z^3, x^2 + y^2 + z^2, (x + y + z)^3) \subset k[x, y, z]$. Use Gröbner bases to show that $x, y, z \in \operatorname{rad} I$ if $\operatorname{ch}(k) \neq 2, 3$.

20. Let $I = (x^3 + y^3 + z^3, xy + xz + yz, xyz) \subset k[x, y, z]$. Use Gröbner bases to show that $x, y, z \in \operatorname{rad} I$.

21. Let $I = (x^4 + y^4 + z^4, x + y + z) \subset k[x, y, z]$.

(a) Use Gröbner bases to show that $xy + xz + yz \in \operatorname{rad} I$ if $\operatorname{ch}(k) \neq 2$ and determine the smallest power of $xy + xz + yz$ contained in I. Show that none of x, y or z is contained in $\operatorname{rad} I$.

(b) If $J = (x^4 + y^4 + z^4, x + y + z, xy + xz + yz)$ show that the reduced Gröbner basis of J relative to the lexicographic ordering $x > y > z$ is $\{x + y + z, y^2 + yz + z^2\}$. Deduce that $k[x, y, z]/J \cong k[y, z]/(y^2 + yz + z^2)$ and that J is radical if $\operatorname{ch}(k) \neq 3$.

(c) If $\operatorname{ch}(k) \neq 2, 3$, show that $\operatorname{rad} I = J$.

(d) If $\operatorname{ch}(k) = 3$, show that $\operatorname{rad} I = (x - y, y - z)$.

(e) If $\operatorname{ch}(k) = 2$, show that $I = (x + y + z)$ is a prime, hence radical, ideal.

22. Let $I = (x^2y + z^3, x + y^3 - z, 2y^4z - yz^2 - z^3) \subset k[x, y, z]$. Use Gröbner bases to show that $x, y, z \in \operatorname{rad} I$ and conclude that $\operatorname{rad} I = (x, y, z)$. Show that x^9, y^7, z^9 are the smallest powers of x, y, z, respectively, lying in I.

23. Let $V = \mathcal{Z}(x^3 - x^2z - y^2z)$ and $W = \mathcal{Z}(x^2 + y^2 - z^2)$ in $\mathbb{C}^3$. Show that $\mathcal{I}(V) = (x^3 - x^2z - y^2z)$ and $\mathcal{I}(W) = (x^2 + y^2 - z^2)$ in $\mathbb{C}[x, y, z]$.

24. Let $V = \mathcal{Z}(x^3 + y^3 + 7z^3) \subset \mathbb{C}^3$. Show that $\mathcal{I}(V) = (x^3 + y^3 + 7z^3)$ in $\mathbb{C}[x, y, z]$.

25. Let $I = (xz + y^2 + z^2, xy - xz + yz - 2z^2)$ and let $K = I + (x^2 - 3y^2 + yz) \subset \mathbb{C}[x, y, z]$.

(a) By Exercise 46 in Section 1, there is an injective $\mathbb{C}$-algebra homomorphism from $\mathbb{C}[x, y, z]/K$ to $\mathbb{C}[u, v]/(u^3 - uv^2 + v^3)$. Use this together with the example preceding Proposition 34 to prove that K is a radical ideal and deduce that $\operatorname{rad} I \subseteq K$.

(b) Show that $\operatorname{rad} I \subseteq (y, z)$.

(c) Show that $K \cap (y, z) = I$ and deduce that I is radical, so that $\mathcal{I}(V) = I$ if $V = \mathcal{Z}(I)$.

(d) Show that $y(x^2 - 3y^2 + yz)$ and $z(x^2 - 3y^2 + yz)$ are elements of I but none of y, z, or $x^2 - 3y^2 + yz$ is contained in I.

26. Let I be an ideal in $k[x_1, \dots, x_n]$. Prove that the following are equivalent (an ideal satisfying any of these conditions is called a *zero-dimensional ideal* because of (d)):

(a) The quotient $k[x_1, \dots, x_n]/I$ has finite dimension as a vector space over k.

(b) $I \cap k[x_i] \neq 0$ for each $i = 1, 2, \dots, n$.

(c) If G is any reduced Gröbner basis for I then for each $i = 1, \dots, n$, there is a $g_i \in G$ with leading term $x_i^{n_i}$ for some $n_i \geq 1$.

(d) The set of common zeros $\mathcal{Z}_{\bar{k}}(I)$ of the polynomials in I in an algebraic closure $\bar{k}$ of k is finite.

[For (a) implies (b) use the injection $k[x_i]/(I \cap k[x_i]) \hookrightarrow k[x_1, \dots, x_n]/I$. For (b) implies (c) note some $LT(g_i)$ divides the leading term of a generator for $I \cap k[x_i]$. For (c) implies (a) use Exercise 37 in Section 9.6. Show (b) implies (d). For (d) implies (b) show the product $m_{a_1,k}(x_i)\dots m_{a_N,k}(x_i)$ of the minimal polynomials of the i^{th} coordinates $a_1, \dots, a_N$ of the points in $\mathcal{Z}_{\bar{k}}(I)$ is a nonzero polynomial in $\mathcal{I}(\mathcal{Z}_{\bar{k}}(I))$ and apply Corollary 33.]

27. Let I be a zero-dimensional ideal in $k[x_1, \dots, x_n]$ and let I' be the ideal generated by I in $\bar{k}[x_1, \dots, x_n]$ where $\bar{k}$ is the algebraic closure of k. Let $\mathcal{Z}(I)$ be the zero set of I in k^n and let $\mathcal{Z}_{\bar{k}}(I)$ be the zero set of I (equivalently, of I') in $\bar{k}^n$.

(a) Prove that $|\mathcal{Z}_{\bar{k}}(I)| = \dim_{\bar{k}} \bar{k}[x_1, \dots, x_n]/\operatorname{rad} I'$. [Show that $\operatorname{rad} I'$ is the product of the maximal ideals corresponding to the points in $V_{\bar{k}}$ and use the Chinese Remainder Theorem.]

(b) Show $|\mathcal{Z}(I)| \leq \dim_k k[x_1, \dots, x_n]/I$. [One approach: use Exercise 43 in Section 1 and observe that $\dim_{\bar{k}} \bar{k}[x_1, \dots, x_n]/\operatorname{rad} I' \leq \dim_{\bar{k}} \bar{k}[x_1, \dots, x_n]/I'$.]

28. Suppose I is a zero-dimensional ideal in $k[x_1, \dots, x_n]$, and suppose $I \cap k[x_i]$ is generated by the nonzero polynomial h_i (cf. Exercise 26). Let r_i be the product of the irreducible factors of h_i (the 'squarefree part' of h_i).

(a) Prove that $I + (r_1, \dots, r_n) \subseteq \operatorname{rad} I$.

(b) *(Radicals of zero-dimensional ideals for perfect fields)* If k is a perfect field, prove that $\operatorname{rad} I = I + (r_1, \dots, r_n)$. [Use induction on n. Write $r_1 = p_1 \dots p_t$ with distinct irreducibles p_i in $k[x_1]$. If $J = I + (r_1, \dots, r_n)$ show that $J = J_1 \cap \dots \cap J_t$ where $J_t = J + (p_t)$. Show for each i that reduction modulo p_i induces an isomorphism $k[x_1, \dots, x_n]/J_i \cong K[x_2, \dots, x_n]/J_i'$ where K is the extension field $k[x]/(p_i)$ and $J_i' \subseteq K[x_2, \dots, x_n]$ is the reduction of the ideal J_i modulo (p_i). Use Exercise 11 of Section 13.5 to show that the image of r_j in $J_i' \cap K[x_j]$ remains a nonzero squarefree polynomial for each $j = 2, \dots, n$ since k is perfect. Conclude by induction that J_i' is a radical ideal. Deduce that J_i is a radical ideal, and finally that J is a radical ideal.]

(c) Find the radicals of $(x^7 + x + y^3, x^4 + y^3 + y)$, $(x^3 - xy^2 + x, x^2y + y^3)$, and $(x^4 + y^3, x^3 - xy + y^2)$ in $\mathbb{Q}[x, y]$ and of $(x^2 + y^2z, x^2y^2 + z^3, y^2 + z^2)$ in $\mathbb{Q}[x, y, z]$.

(d) Let $k = \mathbb{F}_p(t)$. Show that $I = (x^p + t, y^p - t)$ is a zero-dimensional ideal in $k[x, y]$ such that both $I \cap k[x]$ and $I \cap k[y]$ contain nonzero squarefree polynomials, but that I is not a radical ideal (so the result in (b) need not hold if k is not perfect). [Show that $x + y \in \operatorname{rad} I$ but $x + y \notin I$.]

15.4 LOCALIZATION

The idea of "localization at a prime" in a ring is an extremely powerful and pervasive tool in algebra for isolating the behavior of the ideals in a ring. It is an algebraic analogue of the familiar idea of localizing at a point when considering questions of, for example, the differentiability of a function $f(x)$ on the real line. In fact one of the important applications (and also one of the original motivations for the development) of this technique is to translate such "local" properties in the geometry of affine algebraic spaces to corresponding properties of their coordinate rings.

We first consider a very general construction of "rings of fractions." Let D be a multiplicatively closed subset of R containing 1 (i.e., $1 \in D$ and $ab \in D$ if $a, b \in D$). The next result constructs a new ring $D^{-1}R$ which is the "smallest" ring in which the elements of D become units. This generalizes the construction of rings of fractions in Section 7.5 by allowing D to contain zero or zero divisors, and so in this case R need not embed as a subring of $D^{-1}R$.

Theorem 36. Let R be a commutative ring with 1 and let D be a multiplicatively closed subset of R containing 1. Then there is a commutative ring $D^{-1}R$ and a ring homomorphism $\pi : R \to D^{-1}R$ satisfying the following universal property: for any homomorphism $\psi : R \to S$ of commutative rings that sends 1 to 1 such that $\psi(d)$ is a unit in S for every $d \in D$, there is a unique homomorphism $\Psi : D^{-1}R \to S$ such that $\Psi \circ \pi = \psi$.

Proof: The proof is very similar to the proof of Theorem 15 in Section 7.5. In this case we define a relation on $R \times D$ by

$$(r, d) \sim (s, e) \quad \text{if and only if} \quad x(er - ds) = 0 \quad \text{for some } x \in D.$$

This relation is clearly reflexive and symmetric. If $(r, d) \sim (s, e)$ and $(s, e) \sim (t, f)$ then $x(er - ds) = 0$ and $y(fs - et) = 0$ for some $x, y \in D$. Multiplying the first equation by fy and the second by dx and adding gives $exy(fr - dt) = 0$. Since D is closed under multiplication, $(r, d) \sim (t, f)$ and so $\sim$ is transitive.

Let r/d denote the equivalence class of (r, d) under $\sim$ and let $D^{-1}R$ be the set of these equivalence classes. Define addition and multiplication in $D^{-1}R$ by

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \qquad \text{and} \qquad \frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}.$$

It is an exercise to check that these operations are well defined and make $D^{-1}R$ into a commutative ring with $1 = 1/1$. For each $d \in D$, $d/1$ is a unit in $D^{-1}R$ (even in the degenerate case when $D^{-1}R$ is the zero ring).

Finally, define $\pi : R \to D^{-1}R$ by $\pi(r) = r/1$. It follows easily that π is a ring homomorphism. Suppose that $\psi : R \to S$ is a homomorphism of commutative rings that sends 1 to 1 such that $\psi(d)$ is a unit in S for every $d \in D$. Define

$$\Psi : D^{-1}R \to S \qquad \text{by} \qquad \Psi\left(\frac{r}{d}\right) = \psi(r)\psi(d)^{-1}.$$

This map is well defined because if $r/d = s/e$ then $x(er - ds) = 0$ for some $x \in D$. Then $\psi(x)(\psi(er) - \psi(ds)) = 0$ in S, so $\psi(er) - \psi(ds) = 0$ since $\psi(x)$ is a unit in S, and therefore $\psi(r)\psi(d)^{-1} = \psi(s)\psi(e)^{-1}$. It is immediate that Ψ is a ring homomorphism and $\Psi \circ \pi = \psi$.

Finally, Ψ is unique because every element of $D^{-1}R$ can be written as a product $(r/1)(d/1)^{-1}$. The value of Ψ on each element of the form $x/1$ is uniquely determined by ψ, namely $\Psi(x/1) = \Psi(\pi(x)) = \psi(x)$. Since Ψ is a ring homomorphism, its value on u^{-1} for any unit u is uniquely determined by $\Psi(u)$. Thus Ψ is uniquely determined on every element of $D^{-1}R$, completing the proof.

Corollary 37. In the notation of Theorem 36,

(1) $\ker \pi = \{r \in R \mid xr = 0 \text{ for some } x \in D\}$; in particular, $\pi : R \to D^{-1}R$ is an injection if and only if D contains no zero divisors of R, and

(2) $D^{-1}R = 0$ if and only if $0 \in D$, hence if and only if D contains nilpotent elements.

Proof: By definition, we have $\pi(r) = 0$ if and only if $(r, 1) \sim (0, 1)$, i.e., if and only if $xr = 0$ for some $x \in D$, which is (1). For (2), note that $D^{-1}R = 0$ if and only if the 1 of this ring is zero, i.e., $(1, 1) \sim (0, 1)$. This occurs if and only if $x1 = 0$ for some $x \in D$, i.e., if and only if $0 \in D$.

Definition. The ring $D^{-1}R$ is called the *ring of fractions of R with respect to D* or the *localization of R at D*.

Examples

(1) Let R be an integral domain and let $D = R - \{0\}$. Then $D^{-1}R$ is the field of fractions, Q, of R described in Section 7.5. More generally, if D is any multiplicatively closed subset of $R - \{0\}$, then $D^{-1}R$ is the subring of Q consisting of elements r/d with $r \in R$ and $d \in D$.

(2) Let R be any commutative ring with 1 and let f be any element of R. Let D be the multiplicative set $\{f^n \mid n \geq 0\}$ of nonnegative powers of f in R. Define $R_f = D^{-1}R$. Note that $R_f = 0$ if and only if f is nilpotent. If f is not nilpotent, then f becomes a unit in R_f. It is not difficult to see that

$$R_f \cong R[x]/(xf - 1),$$

where $R[x]$ is the polynomial ring in the variable x (cf. the exercises). Note also that R_f and R_{f^n} are naturally isomorphic for any $n \geq 1$ since both f and f^n are units in both rings. If f is a zero divisor then $\pi : R \to R_f$ does not embed R into R_f. For example, let $R = k[x, y]/(xy)$, and take $f = x$. Then x is a unit in R_x and y is mapped to 0 by the first part of the corollary (explicitly: $y = xy/x = 0$ in R_x). In this case $\pi(R) = k[x] \subset R_f = k[x, x^{-1}]$.

(3) (*Localizing at a Prime*) Let P be a prime ideal in any ring R and let $D = R - P$. By definition of a prime ideal D is multiplicatively closed. Passing to the ring $D^{-1}R$ in this case is called *localizing R at P* and the ring $D^{-1}R$ is denoted by R_P. Every element of R not in P becomes a unit in R_P. For example, if $R = \mathbb{Z}$ and $P = (p)$ is a prime ideal, then

$$\mathbb{Z}_{(p)} = \{\frac{a}{b} \in \mathbb{Q} \mid p \nmid b\} \subseteq \mathbb{Q}$$

and every integer b not divisible by p is a unit.

(4) If V is any nonempty set and k is a field, let R be any ring of k-valued functions on V containing the constant functions (for instance, the ring of all continuous real valued functions on the closed interval $[0, 1]$). For any $a \in V$ let M_a be the ideal of functions in R that vanish at a. Then M_a is the kernel of the ring homomorphism from R to the field k given by evaluating each function in R at a. Since R contains the constant functions, evaluation is surjective and so M_a is a maximal (hence also prime) ideal. The localization of R at this prime ideal is then

$$R_{M_a} = \left\{\frac{f}{g} \mid f, g \in R,\ g(a) \neq 0\right\}.$$

Each function in R_{M_a} can then be evaluated at a by $(f/g)(a) = f(a)/g(a)$, and this value does not depend on the choice of representative for the class f/g, so R_{M_a} becomes a ring of k-valued "rational functions" defined at a.

We next consider extensions and contractions of ideals with respect to the map $\pi : R \to D^{-1}R$ in Theorem 36. To ease some of the notation, if I is an ideal of R, let ${}^eI$ denote the extension of I to $D^{-1}R$ (instead of the more cumbersome $D^{-1}R\,\pi(I)$), and if J is an ideal of $D^{-1}R$, let ${}^cJ$ denote the contraction of J to R.

If I is an ideal of R then it is easy to see that every element of ${}^eI$ can be written in the form a/d for some $a \in I$ and $d \in D$, so the extension of I to $D^{-1}R$ is also frequently denoted by $D^{-1}I$.

Proposition 38. In the preceding notation we have

(1) For any ideal J of $D^{-1}R$ we have $J = {}^e({}^cJ)$. In particular, every ideal of $D^{-1}R$ is the extension of some ideal of R, and distinct ideals of $D^{-1}R$ have distinct contractions in R.

(2) For any ideal I of R we have

$$^c({}^eI) = \{r \in R \mid dr \in I \text{ for some } d \in D\}.$$

Also, ${}^eI = D^{-1}R$ if and only if $I \cap D \neq \emptyset$.

(3) Extension and contraction give a bijective correspondence

$$\left\{ \begin{matrix} \text{prime ideals } P \text{ of } R \\ \text{with } P \cap D = \emptyset \end{matrix} \right\} \quad \begin{matrix} \xrightarrow{e} \\ \xleftarrow[c]{} \end{matrix} \quad \left\{ \text{prime ideals of } D^{-1}R \right\}.$$

(4) If R is Noetherian (or Artinian) then $D^{-1}R$ is Noetherian (Artinian, respectively).

Proof: We always have ${}^e({}^cJ) \subseteq J$. For the reverse inclusion let $a/d \in J$. Then $a/1 = d(a/d) \in J$, and so $a \in \pi^{-1}(J) = {}^cJ$. Thus $a/1 \in {}^e({}^cJ)$, so we also have $(a/1)(1/d) = a/d \in {}^e({}^cJ)$, hence $J = {}^e({}^cJ)$. This proves the first statement in (1) and the second statement follows immediately.

Let $I' = \{r \in R \mid dr \in I \text{ for some } d \in D\}$. We first show $I' \subseteq {}^c({}^eI)$. If $r \in I'$ then there is some $d \in D$ such that $dr = a \in I$. Then $r/1 = a/d \in {}^eI$, so $r \in {}^c({}^eI)$. To show the reverse containment ${}^c({}^eI) \subseteq I'$, let $r \in {}^c({}^eI)$ so that $r/1 = a/d$ for some $a \in I$ and $d \in D$. Then $x(dr - a) = 0$ for some $x \in D$, so $xdr = xa \in I$, and because $xd \in D$ it follows that $r \in I'$. This proves the first assertion of (2). Now ${}^eI = D^{-1}R$ if and only if $1/1 \in {}^eI$, if and only if $1 \in {}^c({}^eI) = I'$. The second assertion of (2) then follows from the definition of I'.

To prove (3) observe first that if Q is a prime ideal in $D^{-1}R$, then its preimage under any homomorphism sending 1 to 1 is a prime ideal (cf. Exercise 13, Section 7.4), so c maps prime ideals of $D^{-1}R$ to prime ideals of R disjoint from D. In the reverse direction, let P be a prime ideal of R disjoint from D and let $Q = {}^eP$ and suppose $(a/d_1)(b/d_2) \in Q$. Then $(ab)/(d_1d_2) \in Q$, so $ab/(d_1d_2) = c/d$ for some $c \in P$ and $d \in D$. Then $x(dab - d_1d_2c) = 0$ for some $x \in D$. Since $c \in P$ we have $xdab \in P$, and since P is a prime ideal disjoint from D we have $ab \in P$. Since P is prime, either $a \in P$ or $b \in P$, hence a/d_1 or b/d_2 is in Q. This proves Q is a prime ideal and shows that e maps prime ideals of R disjoint from D to prime ideals of $D^{-1}R$. Finally, it follows immediately from (2) that $P = {}^c({}^eP)$ for every prime ideal of R disjoint from D. Thus c and e are inverse correspondences, hence are bijections between these sets of prime ideals. This establishes (3).

By (1) every ascending (respectively, descending) chain of distinct ideals in $D^{-1}R$ contracts to an ascending (respectively, descending) chain of distinct ideals in R, giving (4) and completing the proof.

Because $1 \in D$, first localizing the ideal I and then contracting that localization as in (2) results in an ideal in R containing I: $I \subseteq {}^c({}^eI)$.

Definition. Suppose R is a commutative ring with 1 and D is a multiplicatively closed subset containing 1. The *saturation* of the ideal I in R with respect to D is the ideal ${}^c({}^eI)$ in R, where contraction and extension are computed with respect to $\pi : R \mapsto D^{-1}R$. If $I = {}^c({}^eI)$ then I is said to be *saturated* with respect to D.

Loosely speaking, (2) of Proposition 38 shows that the saturation of I consists of elements of R that would lie in I if we allowed denominators from D. The ideal is saturated with respect to D if we don't obtain any additional elements even if we allow denominators from D.

We can apply our results on localization to give an algorithm for determining whether an ideal P in the polynomial ring $k[x_1, \dots, x_n]$ with coefficients in the field k is prime. The basic idea is to use the fact that $k[x_1, \dots, x_i] = k[x_1, \dots, x_{i-1}][x_i]$ to consider inductively whether the ideals $P_i = P \cap k[x_1, \dots, x_i]$ are prime.

In general, suppose R is a commutative ring. If P is a prime ideal in $R[x]$ then $P \cap R$ is a prime ideal in R and so $S = R/(P \cap R)$ is an integral domain. Let F denote its quotient field. We then have two natural ring homomorphisms:

$$R[x] \longrightarrow (R/P \cap R)[x] = S[x] \longrightarrow F[x]$$

where the first is the natural projection homomorphism and the second is the natural inclusion induced by $S \subseteq F$. Note that $F[x]$ is the localization of $S[x]$ with respect to the multiplicatively closed set $D = S - \{0\}$. The next proposition shows that the image of P under the first homomorphism is a prime ideal in $S[x]$ that is saturated with respect to D and extends to a prime ideal in $F[x]$, and that, conversely, we can determine whether an ideal is prime in $R[x]$ by these properties.

Proposition 39. Suppose R is a commutative ring with 1 and I is an ideal in $R[x]$. Then I is a prime ideal in $R[x]$ if and only if

- **i.** $J = I \cap R$ is a prime ideal in R, i.e., $S = R/J$ is an integral domain, and
- **ii.** if $\overline{I}$ is the image of I in $S[x]$ then $\overline{I}F[x]$ is a prime ideal in $F[x]$ satisfying $\overline{I}F[x] \cap S[x] = \overline{I}$.

Proof: Suppose I is a prime ideal in $R[x]$, so that $J = I \cap R$ is a prime ideal in R and $S = R/J$ is an integral domain. By Proposition 2 in Chapter 9, the kernel of the reduction homomorphism $R[x] \mapsto S[x] = (R/J)[x]$ is $J[x]$, which is contained in $I[x]$, so we have a ring isomorphism $R[x]/I \cong S[x]/\overline{I}$. Since $R[x]/I$ is an integral domain, it follows that $\overline{I}$ is a prime ideal in the integral domain $S[x]$. The elements of $\overline{I} \cap S$ are the images of the elements in $R \cap I$, so $\overline{I} \cap S = 0$. Since the ring $F[x]$ is the localization of $S[x]$ with respect to the multiplicatively closed set $S - \{0\}$, condition (ii) follows by Proposition 38(3).

Conversely, if I is not prime, then either J is not prime in R or J is prime in R but $\overline{I}$ is not prime in $S[x]$. In the latter case either $\overline{I}F[x]$ is not prime in $F[x]$ or, again by Proposition 38(3), $\overline{I}$ is not saturated. Thus, if I is not prime, either (i) or (ii) fails, completing the proof.

Since $F[x]$ is a Euclidean Domain, the ideal $\overline{I}F[x] = (h(x))$ in Proposition 39 is principal, and is prime if and only if $h(x)$ is either 0 or is irreducible in $F[x]$. Suppose $h(x)$ is an element in I whose image in $S[x]$ has leading coefficient $a \in S$. The next proposition shows that a gives a bound on the denominators necessary for the saturation $\overline{I}F[x] \cap S[x]$ and can be used to compute this saturation.

Proposition 40. Let S be an integral domain with fraction field F and let A be a nonzero ideal in $S[x]$. Suppose $AF[x] = (h(x))$ where $h(x)$ is a polynomial in $S[x]$ with leading coefficient $a \in S$. Let S_a be the localization of S with respect to the powers of a. Then

(1) $AF[x] \cap S[x] = AS_a[x] \cap S[x]$, and

(2) if $\mathcal{A}$ denotes the ideal generated by A and $1 - at$ in the polynomial ring $S[x, t]$, then $AS_a[x] \cap S[x] = \mathcal{A} \cap S[x]$.

Proof: We first show $AF[x] \cap S_a[x] = AS_a[x]$. Since $S_a \subseteq F$, the containment $AS_a[x] \subseteq AF[x] \cap S_a[x]$ is immediate. Suppose now that $f(x) \in AF[x] \cap S_a[x]$. If the leading term of $f(x)$ is sx^N and the leading term of $h(x)$ is ax^m, then since $AF[x] = (h(x))$ we have $N \geq m$. Then the polynomial $f(x) - (s/a)x^{N-m}h(x)$ is again in $AF[x] \cap S_a[x]$ and is of lower degree than $f(x)$. Iterating, we see that $f(x)$ can be written as a polynomial in $S_a[x]$ times $h(x)$, so $f(x) \in AS_a[x]$. Intersecting both sides of $AF[x] \cap S_a[x] = AS_a[x]$ with $S[x]$ gives the first statement in the proposition.

To prove the second statement, suppose first that $f(x) \in \mathcal{A} \cap S[x]$. Then we can write $f(x) = f_1(x,t)b(x) + f_2(x,t)(1 - at)$ for some polynomials $b(x) \in A$ and $f_1, f_2 \in S[x, t]$. Substituting $t = 1/a$ gives $f(x) = f_1(x, 1/a)b(x)$, and since $f_1(x, 1/a) \in S_a[x]$, we obtain $f(x) \in AS_a[x] \cap S[x]$. Conversely, suppose that $f(x) = b(x)g(x) \in S[x]$ where $g(x) \in S_a(x)$ and $b(x) \in A$. If a^N is the largest power of a appearing in the denominators of the coefficients of $g(x)$ then $a^N g(x) \in S[x]$. Writing $f(x) = (at)^N f(x) + (1 - (at)^N) f(x) = b(x)t^N(a^N g(x)) + (1 - (at)^N) f(x)$ we see that $f(x) \in \mathcal{A} \cap S[x]$, giving the reverse containment and completing the proof.

Suppose now that P is an ideal in $k[x_1, \dots, x_n]$. Let P_i for $i = 1, \dots, n$ be the intersection of P with $k[x_1, \dots, x_i]$. We use Propositions 39 and 40 to determine inductively whether $P_1, P_2, \dots, P_n = P$ are prime ideals in their respective polynomial rings.

The ideal P_1 will be prime in the Euclidean Domain $k[x_1]$ if and only if it is 0 or is generated by an irreducible polynomial. Suppose now that $i \geq 2$ and we have already proved that P_{i-1} is a prime ideal in $k[x_1, \dots, x_{i-1}]$, so that the quotient ring $S = k[x_1, \dots, x_{i-1}]/P_{i-1}$ is an integral domain. If F denotes the quotient field of S, then by Proposition 39, P_i is a prime ideal in $k[x_1, \dots, x_i]$ if and only if its image in $(k[x_1, \dots, x_{i-1}]/P_{i-1})[x_i] = S[x_i]$ is a saturated ideal whose extension to the Euclidean Domain $F[x_i]$ is a prime ideal. Suppose $h(x_i) \in S[x_i]$ is a generator for this ideal and a is the leading coefficient of $h(x_i)$. Then $(h(x_i))$ is a prime ideal in $F[x_i]$ if and only if

$h(x_i) = 0$ or $h(x_i)$ is an irreducible polynomial. By Proposition 40, the image of P_i in $S[x_i]$ will be saturated if and only if it equals $\mathcal{A} \cap S[x_i]$ where $\mathcal{A}$ is the ideal generated by P_i and $1 - at$ in $S[x_i, t]$. This latter condition can be checked in $k[x_1, \dots, x_i, t]$: it is equivalent to checking that the intersection of the ideal generated by P_i and $1 - at$ in $k[x_1, \dots, x_i, t]$ with $k[x_1, \dots, x_i]$ is just P_i (cf. Exercise 3).

Combining these observations with our results on Gröbner bases from Chapter 9 we obtain the following algorithm for determining whether the ideal P in $k[x_1, \dots, x_n]$ is prime (or, equivalently, whether the associated affine algebraic set is a variety).

Algorithm for Determining when an Ideal in $k[x_1, \dots, x_n]$ is Prime

(1) Compute the reduced Gröbner basis $G = \{g_1, \dots, g_m\}$ for P with respect to the lexicographic monomial ordering $x_n > \cdots > x_1$.

By Proposition 29 in Section 9.6 the elements of G lying in $k[x_1, \dots, x_i]$ will be the reduced Gröbner basis $\{g_1, \dots, g_{m_i}\}$ for $P_i = P \cap k[x_1, \dots, x_i]$.

(2) Determine whether P_1 is a prime ideal in $k[x_1]$ by checking that $P_1 = 0$ or the nonzero generator of P_1 is irreducible in $k[x_1]$.

For each $i \geq 2$, suppose P_{i-1} has been determined to be a prime ideal in $k[x_1, \dots, x_{i-1}]$ (otherwise, P is not a prime ideal in $k[x_1, \dots, x_n]$). Let $S = k[x_1, \dots, x_{i-1}]/P_{i-1}$ and let F be the fraction field of S. Apply steps (3) and (4) to determine whether P_i is a prime ideal in $k[x_1, \dots, x_i]$.

(3) If $m_i = m_{i-1}$ then P_i maps to the zero ideal in $S[x_i]$, hence is prime. Otherwise the image of P_i in $S[x_i]$ and in $F[x_i]$ is a nonzero ideal, and is generated by the images of $g_{m_{i-1}+1}, \dots, g_{m_i}$. Apply the Euclidean algorithm in $F[x_i]$ to these generators to find an element $h(x_i)$ in P_i whose image in $F[x_i]$ generates the image of P_i in $F[x_i]$. Determine whether $h(x_i)$ is irreducible in $F[x_i]$—if not then P_i and P are not prime ideals.

(Note that after applying the Euclidean algorithm to the generators of the image of P_i in $F[x_i]$ we can multiply by a single element of S to 'clear denominators' in each equation so that all remainders (and in particular the last nonzero remainder $h(x_i)$) will be elements in the image of P_i.)

(4) Let $a \in k[x_1, \dots, x_{i-1}]$ be the leading coefficient of $h(x_i)$ (as a polynomial in x_i). Compute the reduced Gröbner basis in $k[x_1, \dots, x_i, t]$ for the ideal generated by P_i and $1 - at$ with respect to the lexicographic monomial ordering $t > x_i > \cdots > x_1$. Determine whether the elements of this reduced basis that lie in $k[x_1, \dots, x_i]$ are $\{g_1, \dots, g_{m_i}\}$—if so, then P_i is a prime ideal in $k[x_1, \dots, x_i]$ and if not then P_i and P are not prime ideals.

Finally, we note that similar ideas (together with some minor modifications to extend results on Gröbner bases to polynomial rings $R[x_1, \dots, x_n]$ with coefficients in an integral domain R) can be used to provide algorithms for determining when an ideal in, for example, $\mathbb{Z}[x_1, \dots, x_n]$ is prime.

Examples

(1) Consider the ideal $P = (xz - y^2, yz - x^3, z^2 - x^2y)$ in $k[x, y, z]$ for any infinite field k. It follows from Exercise 26 in Section 1 that P is a prime ideal since there is an injection of $k[x, y, z]/P$ into the integral domain $k[\mathbb{A}^1]$ (cf. Exercise 24 in Section 2). Here we prove $P \subset \mathbb{Q}[x, y, z]$ is prime using the ideas in this section. The reduced Gröbner basis for P with respect to the lexicographic monomial ordering $x > y > z$ is $\{x^3 - yz, x^2y - z^2, xy^3 - z^3, xz - y^2, y^5 - z^4\}$. Hence $P_1 = P \cap \mathbb{Q}[z] = (0)$, and $P_2 \cap \mathbb{Q}[y, z] = (y^5 - z^4)$. Since $P_1 = 0$, the ideal P_1 is prime in $\mathbb{Q}[z]$.

We next check P_2 is prime in $\mathbb{Q}[y, z]$, which can be done directly (cf. Exercise 4 or Exercise 14 in Section 9.1). In this case $S = \mathbb{Q}[z]$ and $F = \mathbb{Q}(z)$. The image of P_2 in $F[y]$ is generated by $h(y) = y^5 - z^4$, which is irreducible in $\mathbb{Q}(z)[y]$. The leading coefficient of $h(y)$ is 1, and the reduced Gröbner basis for $(y^5 - z^4, 1 - t)$ in $\mathbb{Q}[y, z, t]$ with respect to the lexicographic monomial ordering $t > y > z$ is $\{y^5 - z^4, 1 - t\}$. The element in the reduced Gröbner basis for P_2 is the only element of this basis lying in $\mathbb{Q}[y, z]$ so P_2 is a prime ideal in $\mathbb{Q}[y, z]$.

We now use the fact that P_2 is prime to prove that P is prime. In this case S is the integral domain $\mathbb{Q}[y, z]/P_2 = \mathbb{Q}[y, z]/(y^5 - z^4)$ with quotient field F given by

$$S = \mathbb{Q}[\bar{z}] + \mathbb{Q}[\bar{z}]\bar{y} + \mathbb{Q}[\bar{z}]\bar{y}^2 + \mathbb{Q}[\bar{z}]\bar{y}^3 + \mathbb{Q}[\bar{z}]\bar{y}^4$$
$$F = \mathbb{Q}(\bar{z}) + \mathbb{Q}(\bar{z})\bar{y} + \mathbb{Q}(\bar{z})\bar{y}^2 + \mathbb{Q}(\bar{z})\bar{y}^3 + \mathbb{Q}(\bar{z})\bar{y}^4$$

where $\bar{y}^5 = \bar{z}^4$. The image of P in $S[x]$ is the ideal $\overline{P}$ generated by the elements $g_1 = x^3 - \bar{y}\bar{z}$, $g_2 = \bar{y}x^2 - \bar{z}^2$, $g_3 = \bar{y}^3x - \bar{z}^3$, $g_4 = \bar{z}x - \bar{y}^2$, and $\bar{y}^5 - \bar{z}^4 = 0$.

The greatest common divisor in $F[x]$ of g_1, g_2, g_3, g_4 generating the image of P in $F[x]$ is the irreducible polynomial $x - \bar{y}^2/\bar{z}$. The polynomial $h(x) = zx - y^2$ in P has image generating the same ideal in $F[x]$, so we may take $a = z$ in (4) of the algorithm. The reduced Gröbner basis for $(xz - y^2, yz - x^3, z^2 - x^2y, 1 - zt)$ with respect to the lexicographic monomial ordering $t > x > y > z$ consists of the reduced Gröbner basis for P together with the elements $ty^2 - x$ and $tz - 1$ involving t, so P is a prime ideal in $\mathbb{Q}[x, y, z]$.

(2) Consider the ideal $P = (xz - y^3, xy - z^2)$ in $\mathbb{Q}[x, y, z]$, with reduced Gröbner basis for the lexicographic monomial ordering $x > y > z$ given by $\{xy - z^2, xz - y^3, y^4 - z^3\}$. Here $P_1 = 0$ and $P_2 = P \cap \mathbb{Q}[y, z] = (y^4 - z^3)$ are prime ideals as in Example 1. In this case $S = \mathbb{Q}[y, z]/P_2$ is given by

$$S = \mathbb{Q}[\bar{z}] + \mathbb{Q}[\bar{z}]\bar{y} + \mathbb{Q}[\bar{z}]\bar{y}^2 + \mathbb{Q}[\bar{z}]\bar{y}^3$$

with $\bar{y}^4 = \bar{z}^3$, with quotient field F similar to the previous example, and $\overline{P} = (g_1, g_2)$ in $S[x]$ where $g_1 = \bar{y}x - \bar{z}^2$ and $g_2 = \bar{z}x - \bar{y}^3$. The extension of $\overline{P}$ to $F[x]$ is generated by the irreducible polynomial $\bar{y}x - \bar{z}^2$, and $h(x) = yx - z^2$ is an element of P having the same image in $F[x]$, with leading coefficient $a = y$. The reduced Gröbner basis for the ideal $(xz - y^3, xy - z^2, 1 - yt)$ in $\mathbb{Q}[x, y, z, t]$ using the lexicographic ordering $t > x > y > z$ is $\{x^2 - y^2z, xy - z^2, xz - y^3, y^4 - z^3, ty - 1, tz^2 - x\}$, containing the element $x^2 - y^2z$ not in the reduced Gröbner basis for P, so P is *not* a prime ideal in $\mathbb{Q}[x, y, z]$. This computation not only shows P is not a prime ideal, it does so by explicitly showing the image of P in $S[x]$ is not saturated using the localization S_a. The computation of $a = y$ allows us to find an explicit pair of elements not in P whose product is in P: $f = x^2 - y^2z \notin P$ and $y \notin P$, but some power of y times f lies in P. In this case a quick computation verifies that $yf \in P$.

Localizations of Modules

Suppose now that M is an R-module and D is a multiplicatively closed subset of R containing 1 as above. Then the ideas used in the construction of $D^{-1}R$ can be used to construct a $D^{-1}R$-module $D^{-1}M$ from M in a similar fashion, as follows. Define the relation on $D \times M$ by

$$(d, m) \sim (e, n) \quad \text{if and only if} \quad x(dn - em) = 0 \quad \text{for some } x \in D,$$

which is easily checked to be an equivalence relation. Let m/d denote the equivalence class of (d, m) and let $D^{-1}M$ denote the set of equivalence classes. It is then straightforward to verify that the operations

$$\frac{m}{d} + \frac{n}{e} = \frac{em + dn}{de} \qquad \text{and} \qquad \left(\frac{r}{d}\right)\left(\frac{m}{e}\right) = \frac{rm}{de}$$

are well defined and give $D^{-1}M$ the structure of a $D^{-1}R$-module.

Definition. The $D^{-1}R$-module $D^{-1}M$ is called the *module of fractions of M with respect to D* or the *localization of M at D*.

Note that the localization $D^{-1}M$ is also an R-module (since each $r \in R$ acts by $r/1$ on $D^{-1}M$), and there is an R-module homomorphism

$$\pi : M \to D^{-1}M \quad \text{defined by} \quad \pi(m) = \frac{m}{1}.$$

It follows directly from the definition of the equivalence relation that

$$\ker \pi = \{m \in M \mid dm = 0 \text{ for some } d \in D\}.$$

The homomorphism π has a universal property analogous to that in Theorem 36. Suppose N is an R-module with the property that left multiplication on N by d is a bijection of N for every $d \in D$. If $\psi : M \to N$ is any R-module homomorphism then there is a unique R-module homomorphism $\Psi : D^{-1}M \to N$ such that $\Psi \circ \pi = \psi$.

If M and N are R-modules and $\varphi : M \to N$ is an R-module homomorphism, then for any multiplicative set D in R it is easy to check that there is an induced $D^{-1}R$-module homomorphism from $D^{-1}M$ to $D^{-1}N$ defined by mapping m/d to $\varphi(m)/d$.

The next result shows that the localization of M at D is related to the tensor product.

Proposition 41. Let D be a multiplicatively closed subset of R containing 1 and let M be an R-module. Then $D^{-1}M \cong D^{-1}R \otimes_R M$ as $D^{-1}R$-modules, i.e., $D^{-1}M$ is the $D^{-1}R$-module obtained by extension of scalars from the R-module M.

Proof: The map from $D^{-1}R \times M$ to $D^{-1}M$ defined by mapping $(r/d, m)$ to rm/d is well defined and R-balanced, so induces a homomorphism from $D^{-1}R \otimes_R M$ to $D^{-1}M$. The map sending m/d to $(1/d) \otimes m$ gives a well defined inverse homomorphism (if $m/d = m'/d'$ in $D^{-1}M$ then $x(d'm - dm') = 0$ for some $x \in D$, and then $(1/d) \otimes m$ can be written as $(1/xd'd) \otimes (xd'm) = (1/xd'd) \otimes (xdm') = (1/d') \otimes m'$). Hence $D^{-1}M$ is isomorphic to $D^{-1}R \otimes_R M$ as an R-module since these inverse isomorphisms are also $D^{-1}R$-module homomorphisms.

Localizing a ring R or an R-module M at D behaves very well with respect to algebraic operations on rings and modules, as the following proposition shows:

Proposition 42. Let R be a commutative ring with 1 and let $D^{-1}R$ be its localization with respect to the multiplicatively closed subset D of R containing 1.

(1) Localization commutes with finite sums and intersections of ideals: If I and J are ideals of R, then
$$D^{-1}(I+J) = D^{-1}(I) + D^{-1}(J) \quad \text{and} \quad D^{-1}(\dot{I} \cap J) = D^{-1}(I) \cap D^{-1}(J).$$
Localization commutes with quotients:
$$D^{-1}R \,/\, D^{-1}I \cong D^{-1}(R/I),$$
(where the localization on the right is with respect to the image of D in the quotient R/I).

(2) Localization commutes with taking radicals: If N is the nilradical of R, then $D^{-1}N$ is the nilradical of $D^{-1}R$. If I is an ideal in R, then $\operatorname{rad}(D^{-1}I)$ is $D^{-1}(\operatorname{rad} I)$.

(3) Primary ideals correspond to primary ideals in the correspondence (3) of Proposition 38. More precisely, suppose Q is a P-primary ideal in R. If $D \cap P \neq \emptyset$ then $D^{-1}Q = D^{-1}R$. If $D \cap P = \emptyset$ then $D^{-1}P$ is a prime ideal, the extension $D^{-1}Q$ of Q is a $D^{-1}P$-primary ideal in $D^{-1}R$, and the contraction back to R of $D^{-1}Q$ is Q.

(4) Localization commutes with finite sums, intersections and quotients of modules: If L and N are submodules of the R-module M, then
(a) $D^{-1}(L+N) = D^{-1}L + D^{-1}N$ and $D^{-1}(L \cap N) = D^{-1}L \cap D^{-1}N$,
(b) $D^{-1}N$ is a submodule of $D^{-1}M$ and $D^{-1}M \,/\, D^{-1}N = D^{-1}(M/N)$.

(5) Localization commutes with finite direct sums of modules: If M and N are R-modules, then $D^{-1}(M \oplus N) \cong D^{-1}M \oplus D^{-1}N$.

(6) Localization is exact (i.e., $D^{-1}R$ is a flat R-module): If $0 \to L \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0$ is a short exact sequence of R-modules, then the induced sequence $0 \to D^{-1}L \xrightarrow{\psi'} D^{-1}M \xrightarrow{\varphi'} D^{-1}N \to 0$ of $D^{-1}R$-modules is also exact.

Proof: We first prove (6). Suppose that $0 \to L \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0$ is a short exact sequence of R-modules. Every element of $D^{-1}N$ is of the form n/d for some $n \in N$ and $d \in D$. Since φ is surjective, $n = \varphi(m)$ for some $m \in M$, so $\varphi'(m/d) = \varphi(m)/d = n/d$ and $\varphi' : D^{-1}M \to D^{-1}N$ is surjective. If m/d is in the kernel of φ' then $d_1\varphi(m) = 0$ for some $d_1 \in D$. Then $\varphi(d_1 m) = 0$ implies $d_1 m = \psi(l)$ for some $l \in L$ by the exactness of the original sequence at M, so $m/d = d_1 m/(d_1 d) = \psi(l)/(d_1 d) = \psi'(l/(d_1 d))$ and $\ker(\varphi') \subseteq \operatorname{image}(\psi')$. If $\psi(l)/d \in \operatorname{image}(\psi')$ then $\varphi'(\psi(l)/d) = \varphi(\psi(l))/d = 0$, which shows the reverse inclusion $\operatorname{image}(\psi') \subseteq \ker(\varphi')$, and we have exactness of the induced sequence at $D^{-1}M$. Finally, suppose $\psi'(l/d) = 0$. Then $d_2\psi(l) = 0$ for some $d_2 \in D$, i.e., $\psi(d_2 l) = 0$, so $d_2 l = 0$ by the injectivity of ψ. Hence $l/d = d_2 l/(d_2 d) = 0$ and ψ' is injective. This proves that the sequence $0 \to D^{-1}L \xrightarrow{\psi'} D^{-1}M \xrightarrow{\varphi'} D^{-1}N \to 0$ is exact.

To prove the first statement in (1), note that $(i+j)/d = i/d + j/d$ for $i \in I$, $j \in J$ and $d \in D$ shows $D^{-1}(I+J) \subseteq D^{-1}(I) + D^{-1}(J)$; and $i/d_1 + j/d_2 = (d_2 i + d_1 j)/(d_1 d_2)$ for $i \in I$, $j \in J$ and $d_1, d_2 \in D$ shows $D^{-1}(I) + D^{-1}(J) \subseteq D^{-1}(I+J)$. For the second statement, the inclusion $D^{-1}(I \cap J) \subseteq D^{-1}(I) \cap D^{-1}(J)$ is immediate. If

$a/d \in D^{-1}(I) \cap D^{-1}(J)$ then $d_1 a \in I$ and $d_2 a \in J$ for some $d_1, d_2 \in D$. Then $d_1 d_2 a \in I \cap J$ and $a/d = (d_1 d_2 a)/(d_1 d_2 d)$ gives the inclusion $D^{-1}(I) \cap D^{-1}(J) \subseteq D^{-1}(I \cap J)$. The last statement in (1) follows by applying (6) to the exact sequence $0 \to I \xrightarrow{\psi} R \xrightarrow{\varphi} R/I \to 0$.

To prove (2), suppose first that $a \in \operatorname{rad} I$, so that $a^n \in I$ for some $n \geq 1$. Then $(a/d)^n = a^n/d^n \in D^{-1}I$ so $D^{-1}(\operatorname{rad} I) \subseteq \operatorname{rad}(D^{-1}I)$. Conversely, if $a/d \in \operatorname{rad}(D^{-1}I)$ then $(a/d)^n \in D^{-1}I$ for some $n \geq 1$, i.e., $d_1 a^n \in I$ for some $d_1 \in D$. Hence $(d_1 a)^n = d_1^{n-1}(d_1 a^n) \in I$, so $d_1 a \in \operatorname{rad} I$ and then $a/d = d_1 a/(d_1 d) \in D^{-1}(\operatorname{rad} I)$ shows that $\operatorname{rad}(D^{-1}I) \subseteq D^{-1}(\operatorname{rad} I)$. This proves the second statement in (2), and the first statement follows by applying this to the ideal $I = (0)$.

For (3), note first that $D \cap P = \emptyset$ if and only if $D \cap Q = \emptyset$ (one inclusion is obvious and the other follows since $d \in D \cap P$ implies $d^n \in D \cap Q$ for some n). The statement for $D \cap P \neq \emptyset$ and the fact that $D^{-1}P$ is a prime ideal for $D \cap P = \emptyset$ were proved in Proposition 38. To see that $D^{-1}Q$ is a primary ideal in $D^{-1}R$, suppose that $(a/d_1)(b/d_2) \in D^{-1}Q$ and $a/d_1 \notin D^{-1}Q$. Then there is some element $d \in D$ so that $dab \in Q$, and since $a \notin Q$ and Q is primary, we have $(db)^n \in Q$ for some $n \geq 1$. Then $(b/d_2)^n = d^n b^n/(d^n d_2^n) \in D^{-1}Q$, so that $D^{-1}Q$ is primary. The radical of $D^{-1}Q$ is $D^{-1}P$ by (2). Finally, by (2) of Proposition 38, the contraction of $D^{-1}Q$ is an ideal of R containing Q and consists precisely of the elements $r \in R$ with $dr \in Q$ for some $d \in D$. Since Q is P-primary, the definition of primary implies that if $dr \in Q$ and $d \notin P$, then $r \in Q$, hence the contraction of $D^{-1}Q$ is Q.

The proof of (4) is essentially the same as the proof of (1) and is left as an exercise.

It is easy to see that if the exact sequence $0 \to L \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0$ of R-modules splits, then the exact sequence $0 \to D^{-1}L \xrightarrow{\psi'} D^{-1}M \xrightarrow{\varphi'} D^{-1}N \to 0$ of $D^{-1}R$-modules also splits, which gives (5).

Proposition 38 shows that localizing at the multiplicatively closed set D emphasizes the ideals of R not containing any elements of D since the other ideals of R become trivial when extended to $D^{-1}R$. The following proposition provides a more precise statement in terms of the effect of localization on primary decomposition of ideals.

Proposition 43. Let R be a Noetherian ring and let

$$I = Q_1 \cap \cdots \cap Q_m$$

be a minimal primary decomposition of the proper ideal I, where Q_i is a P_i-primary ideal. Suppose D is a multiplicatively closed set of R containing 1 and the primary ideals $Q_1, \ldots, Q_m$ are numbered so that $D \cap P_i = \emptyset$ for $1 \leq i \leq t$ and $D \cap P_i \neq \emptyset$ for $t + 1 \leq i \leq m$. Then

$$D^{-1}I = D^{-1}Q_1 \cap \cdots \cap D^{-1}Q_t$$

is a minimal primary decomposition of $D^{-1}I$ in $D^{-1}R$ and $D^{-1}Q_i$ is a $D^{-1}P_i$-primary ideal. Further, the contraction of $D^{-1}Q_i$ back to R is Q_i for $1 \leq i \leq t$ and

$$^{c}(D^{-1}I) = Q_1 \cap \cdots \cap Q_t$$

is a minimal primary decomposition of the contraction of $D^{-1}I$ back to R.

Proof: By (3) of Proposition 42, $D^{-1}Q_i = D^{-1}R$ for $t+1 \leq i \leq m$, and $D^{-1}Q_i$ is a $D^{-1}P_i$-primary ideal with pullback Q_i for $1 \leq i \leq t$. By (1) of the same proposition, $D^{-1}I = D^{-1}Q_1 \cap \cdots \cap D^{-1}Q_t$, and (3) shows that this is a primary decomposition. Contracting to R shows that ${}^c(D^{-1}I) = Q_1 \cap \cdots \cap Q_t$, which also implies that the decompositions are minimal.

In particular we can finish the proof of Theorem 21:

Corollary 44. The primary ideals belonging to the isolated primes in a minimal primary decomposition of I are uniquely defined by I.

Proof: Let P be a minimal element in the set $\{P_1, \ldots, P_m\}$ of primes belonging to I, and take $D = R - P$ in Proposition 43. Then $D \cap P_i = \emptyset$ only for $P = P_i$, so the contraction of the localization of I at D is precisely the primary ideal Q belonging to the minimal prime P. Since the prime ideals $\{P_1, \ldots, P_m\}$ of primes belonging to I are uniquely determined by I, it follows that the primary ideals Q belonging to the isolated primes of I are also uniquely determined by I.

The effect of isolating in on certain prime ideals by localization is particularly precise in the case of localizing at a prime P (considered in Example 3 following Corollary 37 above). We first recall the definition of an important type of ring (cf. Exercises 37–39 in Section 7.4).

Definition. A commutative ring with 1 that has a unique maximal ideal is called a *local ring*.

Proposition 45. Let R be a commutative ring with 1. Then the following are equivalent:

(1) R is a local ring with unique maximal ideal M

(2) if M is the set of elements of R that are not units, then M is an ideal

(3) there is a maximal ideal M of R such that every element $1+m$ with $m \in M$ is a unit in R.

Proof: If $a \in R$ then the ideal (a) is either R, in which case a is a unit, or is a proper ideal, in which case (a) is contained in a maximal ideal (Proposition 11 of Section 7.4). It follows that if R is a local ring and M is its unique maximal ideal then every $a \notin M$ is a unit, so M consists precisely of the set of nonunits in R, showing that (1) implies (2). It also follows that if the set M of nonunits in R is an ideal then this ideal must be the unique maximal ideal in R, so that (2) implies (1).

Suppose now that (3) is satisfied. If a is an element of R not contained in the maximal ideal M, then $(a) + M = R$, so that $ab + m = 1$ for some $b \in R$ and $m \in M$. Then $ab = 1 - m$ is a unit by assumption, so a is also a unit. This shows that M is the unique maximal ideal in R, so (3) implies (1). Conversely, if R is a local ring, then $1 + m \notin M$ for any $m \in M$, so $1 + m$ is a unit, so (1) implies (3).

Proposition 46. For any commutative ring R with 1, let R_P be the localization of R at the prime ideal P and let ${}^eP$ be the extension of P to R_P.

(1) The ring R_P is a local ring with unique maximal ideal ${}^eP$. The contraction of ${}^eP$ to R is P, i.e., ${}^c({}^eP) = P$, and the map from R to R_P induces an injection of the integral domain R/P into $R_P/{}^eP$. The quotient $R_P/{}^eP$ is a field and is isomorphic to the fraction field of the integral domain R/P.

(2) If R is an integral domain, then R_P is an integral domain. The ring R injects into the local ring R_P, and, identifying R with its image in R_P, the unique maximal ideal of R_P is PR_P.

(3) The prime ideals in R_P are in bijective correspondence with the prime ideals of R contained in P.

(4) If P is a minimal nonzero prime ideal of R then R_P has a unique nonzero prime ideal.

(5) If $P = M$ is a maximal ideal and I is any M-primary ideal of R then $R_M/{}^eI \cong R/I$. In particular, $R_M/{}^eM \cong R/M$ and $({}^eM)/({}^eM)^n \cong M/M^n$ for all $n \geq 1$.

Proof: If P' is a prime ideal of R, then $P' \cap (R - P) = \emptyset$ if and only if $P' \subseteq P$, so (3) is immediate from (3) in Proposition 38, and (4) follows. Since ${}^eP \neq R_P$ by (2) of Proposition 38, it follows from (3) that R_P is a local ring with unique maximal ideal ${}^eP$, which proves the first statement in (1).

By Proposition 38(2) the contraction ${}^c({}^eP)$ is the set $\{r \in R \mid dr \in P \text{ for some } d \in R - P\}$, and since P is prime, $dr \in P$ with $d \notin P$ implies $r \in P$. This shows that ${}^c({}^eP) = P$, which is the second statement in (1).

The kernel of the map from R to $R_P/{}^eP$ is ${}^c({}^eP) = P$, so the induced map from R/P into $R_P/{}^eP$ is injective. The quotient $R_P/{}^eP$ is a field by the first part of (1), so there is an induced homomorphism from the fraction field of the integral domain R/P into $R_P/{}^eP$. The universal property of the localization R_P shows there is an inverse homomorphism from $R_P/{}^eP$ to the fraction field of R/P (since every element of R not in P maps to a unit in R/P). It follows that $R_P/{}^eP$ is isomorphic to the fraction field of R/P.

If R is an integral domain, then $R - P$ has no zero divisors, so R injects into R_P by Corollary 37; if R is identified with its image in R_P then ${}^eP = PR_P$, so (2) follows.

To prove (5), by Proposition 42(1) we may pass to the quotient R/I and so reduce to the case $I = 0$. In this case the maximal ideal $P = M$ in R is the nilradical of R, hence is the unique maximal ideal of R. By Proposition 45 every element of $R - M$ is a unit, so $R_P = R$, and each of the statements in (5) follows immediately, completing the proof of the proposition.

Example

The results of (5) of the proposition are not true in general if P is a prime ideal that is not maximal. For example, $P = (0)$ in $R = \mathbb{Z}$ has $R/P = \mathbb{Z}$ and $R_P/PR_P = \mathbb{Q}$; in this case $(PR_P)/(PR_P)^n \cong P/P^n = 0$ for all $n \geq 1$ (cf. the exercises).

Definition. Let M be an R-module, let P be a prime ideal of R and set $D = R - P$. The R_P-module $D^{-1}M$ is called the *localization of M at P*, and is denoted by M_P.

By Proposition 41, M_P can also be identified with the tensor product $R_P \otimes_R M$. When R is an integral domain and $P = (0)$, then $M_{(0)}$ is a module over the field of fractions F of R, i.e., is a vector space over F.

The element $m/1$ is zero in M_P if and only if $rm = 0$ for some $r \in R - P$, so localizing at P annihilates the P'-torsion elements of M for primes P' not contained in P. In particular, *localizing at (0) over an integral domain annihilates the torsion subgroup of M.*

Definition. If R is an integral domain, then the *rank* of the R-module M is the dimension of the localization $M_{(0)}$ as a vector space over the field of fractions of R.

It is easy to see that this definition of rank agrees with the notion of rank introduced in Chapter 12.

Example

Let $R = \mathbb{Z}$ and let $\mathbb{Z}_{(p)}$ be the localization of $\mathbb{Z}$ at the nonzero prime ideal (p). Any abelian group M is a $\mathbb{Z}$-module so we may localize M at (p) by forming $M_{(p)}$. This abelian group is the same as the quotient of M with respect to the subgroup of elements whose order is finite and not divisible by p. If M is a finite (or, more generally, torsion) abelian group, then $M_{(p)}$ is a p-group, and is the Sylow p-subgroup or p-primary component of M. The localization $M_{(0)}$ of M at (0) is the trivial group. For a specific example, let $M = \mathbb{Z}/6\mathbb{Z}$ be the cyclic group of order 6, considered as a $\mathbb{Z}$-module. Then the localization of M at $p = 2$ is $\mathbb{Z}/2\mathbb{Z}$, at $p = 3$ is $\mathbb{Z}/3\mathbb{Z}$, and reduces to 0 at all other prime ideals of $\mathbb{Z}$.

Localization of a module M at a prime P in general produces a simpler module M_P whose properties are easier to determine. It is then of interest to translate these "local" properties of M_P back into "global" information about the module M itself. For example, the most basic question of whether a module M is 0 can be answered locally:

Proposition 47. Let M be an R-module. Then the following are equivalent:

(1) $M = 0$,
(2) $M_P = 0$ for all prime ideals P of R, and
(3) $M_{\mathfrak{m}} = 0$ for all maximal ideals $\mathfrak{m}$ of R.

Proof: The implications (1) implies (2) implies (3) are obvious, so it remains to prove that (3) implies (1). Suppose m is a nonzero element in M, and consider the annihilator I of m in R, i.e., the ideal of elements $r \in R$ with $rm = 0$. Since m is nonzero I is a proper ideal in R. Let $\mathfrak{m}$ be a maximal ideal of R containing I and consider the element $m/1$ in the corresponding localization $M_{\mathfrak{m}}$ of M. If this element were 0, then $rm = 0$ for some $r \in R - \mathfrak{m}$. But then r would be an element in I not contained in $\mathfrak{m}$, a contradiction. Hence $M_{\mathfrak{m}} \neq 0$, which proves that (3) implies (1).

It is not in general true that a property shared by all of the localizations of a module M is also shared by M. For example, all of the localizations of a ring R can be integral domains without R itself being an integral domain (for example, $\mathbb{Z}/6\mathbb{Z}$ above). Nevertheless, a great deal of information *can* be ascertained from studying the various possible localizations, and this is what makes this technique so useful. If R is an integral domain, for example, then each of the localizations R_P can be considered as a subring of the fraction field F of R that contains R; the next proposition shows that the elements of R are the only elements of F contained in every localization.

Proposition 48. Let R be an integral domain. Then R is the intersection of the localizations of R: $R = \cap_P R_P$. In fact, $R = \cap_{\mathfrak{m}} R_{\mathfrak{m}}$ is the intersection of the localizations of R at the maximal ideals $\mathfrak{m}$ of R.

Proof: As mentioned, $R \subseteq \cap_{\mathfrak{m}} R_{\mathfrak{m}}$. Suppose now that a is an element of the fraction field F of R that is contained in $R_{\mathfrak{m}}$ for every maximal ideal $\mathfrak{m}$ of R, and consider

$$I_a = \{d \in R \mid da \in R\}.$$

It is easy to check that I is an ideal of R, and that $a \in R$ if and only if $1 \in I_a$, i.e., $I_a = R$. Suppose that $I_a \neq R$. Then there is a maximal ideal $\mathfrak{m}$ containing I_a, and since $a \in R_{\mathfrak{m}}$ we have $a = r/d$ for some $r \in R$ and $d \in R - \mathfrak{m}$. But then $d \in I_a$ and $d \notin \mathfrak{m}$, a contradiction. Hence $a \in R$, so $\cap_{\mathfrak{m}} R_{\mathfrak{m}} \subseteq R$, and we have proved the second assertion in the proposition. The first is then immediate.

Another important property of a ring R that can be detected locally is normality:

Proposition 49. Let R be an integral domain. Then the following are equivalent:

(1) R is normal, i.e., R is integrally closed (in its field of fractions)

(2) R_P is normal for all prime ideals P of R

(3) $R_{\mathfrak{m}}$ is normal for all maximal ideals $\mathfrak{m}$ of R.

Proof: Let F be the field of fractions of R, so all of the various localizations of R may be considered as subrings of F.

Assume first that R is integrally closed and suppose $y \in F$ is integral over R_P. Then y is a root of a monic polynomial of degree n with coefficients of the form a_i/d_i for some $d_i \notin P$. The element $y' = y(d_0 d_1 \cdots d_{n-1})^n$ is then a root of a monic polynomial of degree n with coefficients from R, i.e., y' is integral over R. Since R is assumed normal, this implies $y' \in R$, and so $y = y'/(d_0 \cdots d_{n-1}) \in R_P$, which proves that (1) implies (2). The implication (2) implies (3) is trivial. Suppose now that $R_{\mathfrak{m}}$ is normal for all maximal ideals $\mathfrak{m}$ of R and let y be an element of F that is integral over R. Since $R \subseteq R_{\mathfrak{m}}$, y is in particular also integral over $R_{\mathfrak{m}}$ and so $y \in R_{\mathfrak{m}}$ for every maximal ideal by assumption. Then $y \in R$ by the previous proposition, which proves that (3) implies (1).

We now may easily prove the first part of the Going–up Theorem (cf. Section 3) that was used in the proof of Corollary 27.

Corollary 50. Let R be a subring of the commutative ring S with $1 \in R$, and assume that S is integral over R. If P is a prime ideal in R, then there is a prime ideal Q of S with $P = Q \cap R$.

Proof: Let $D = R - P$ so that D is a multiplicatively closed subset of both R and S. Then the following diagram commutes:

$$\begin{tikzcd}
R \arrow[r, "\pi"] \arrow[d, "\iota"'] & D^{-1}R = R_P \arrow[d, "\iota"] \\
S \arrow[r, "\pi"] & D^{-1}S
\end{tikzcd}$$

where the vertical maps are inclusions. It is easy to see that $D^{-1}S$ is integral over R_P (Exercise 20). Let $\mathfrak{m}$ be any maximal ideal of $D^{-1}S$. Then $\mathfrak{m} \cap R_P$ is a maximal ideal in R_P by the second statement in Theorem 26(2) (note that the first part of Theorem 26(2) was not used in the proof of the second statement). By Proposition 38(1), $\mathfrak{m} \cap R_P$ is the extension of P to the local ring R_P, and the contraction of this ideal to R is just P. Put another way, the preimage of $\mathfrak{m}$ by the maps along the top and right of the diagram above is P. If $Q \subset S$ denotes the preimage of $\mathfrak{m}$ by the map along the bottom of the diagram, then Q is a prime ideal by Proposition 38(3). Since $Q \cap R$ is the pullback of Q by the map along the left of the diagram above, the commutativity of the diagram shows that $Q \cap R = P$.

Local Rings of Affine Algebraic Varieties

For the remainder of this section, let k be an algebraically closed field and let V be an affine variety over k with coordinate ring $k[V]$. Then $k[V]$ is an integral domain, so we may form its field of fractions:

$$k(V) = \{f/g \mid f, g \in k[V],\ g \neq 0\}.$$

The elements of $k(V)$ are called *rational functions* on V and $k(V)$ is called the *field of rational functions* on V. When $k[V]$ is a Unique Factorization Domain there is an essentially unique representative for f/g that is in "lowest terms," but in general each fraction $f/g \in k(V)$ has many representations as a ratio of two elements of $k[V]$. Since $k[V]$ is an integral domain, $f/g = f_1/g_1$ if and only if $fg_1 = f_1g$.

The elements of $k[V]$ can be considered as k-valued functions on V, and if the denominator doesn't vanish the same is true for an element of $k(V)$ (which helps to explain the terminology for this field). Since the same element of $k(V)$ may be written in the form f/g in several ways, we make the following definition:

Definition. We say f/g is *regular at* v or *defined at the point* $v \in V$ if there is some $f_1, g_1 \in k[V]$ with $f/g = f_1/g_1$ and $g_1(v) \neq 0$.

If f_2, g_2 is another such pair with $g_2(v) \neq 0$, then $f_1(v)/g_1(v) = f_2(v)/g_2(v)$ as elements of k, so whenever f/g is regular at v there is a well defined way of specifying its value in k at v.

Example

The variety $V = \mathcal{Z}(xz - yw)$ in $\mathbb{A}^4$ has coordinate ring $k[V] = k[x, y, z, w]/(xz - yw)$. Consider the element $f = \bar{x}/\bar{y}$ in the quotient field $k(V)$ of $k[V]$. Since $\bar{x}\bar{z} = \bar{y}\bar{w}$ in $k[V]$, the element f can also be written as $\bar{w}/\bar{z}$. From the first expression for f it follows that f

is regular at all points of V where $\bar{y} \neq 0$, and from the second expression it follows that f is regular at all points of V where $\bar{z} \neq 0$. It is not too difficult to show that these are all the points of V where f is regular. Furthermore, there is no single expression $f = a/b$ for f with $a, b \in k[V]$ such that $b(v) \neq 0$ for every v where f is regular (cf. Exercise 25).

If $f/g \in k(V)$ is regular at the point v, say $f/g = f_1/g_1$ with $g_1(v) \neq 0$, then f/g is also regular at all the points v in the Zariski open neighborhood V_{g_1} of v where $g_1 \neq 0$. As a k-valued function on V this means that if f/g is defined at v, then it is also defined in a (Zariski open) neighborhood of v. Since any nonempty open set of an affine variety is Zariski dense (cf. Exercise 11 in Section 2), we see that every rational function on V is defined at a dense set of points in V (so "almost everywhere" in a suitable sense). Also, each pair f_1/g_1 and f_2/g_2 representing f/g agree as functions on the open neighborhood $V_{g_1} \cap V_{g_2}$ of v, but the "size" of this neighborhood depends on g_1 and g_2 — there is in general not a common open neighborhood of v where *all* representatives of f/g with nonzero denominator at v are simultaneously defined.

If v is a fixed point in V, then a rational function f/g is regular at v if and only if $f/g = f_1/g_1$ for some $f_1, g_1 \in k[V]$ with $g_1 \notin \mathcal{I}(v)$, the ideal of functions on V that are zero at v. This means that the set of rational functions that are defined at v is the same as the localization of $k[V]$ at the maximal ideal $\mathcal{I}(v)$:

Definition. For each point $v \in V$ the collection of rational functions on V that are defined at v,

$$\mathcal{O}_{v,V} = \{f/g \in k(V) \mid f/g \text{ is regular at } v\},$$

is called the *local ring of V at v*. Equivalently, the local ring of V at v is the localization of $k[V]$ at the maximal ideal $\mathcal{I}(v)$.

In particular, $\mathcal{O}_{v,V}$ is a local ring with unique maximal ideal $\mathfrak{m}_{v,V}$, where

$$\mathfrak{m}_{v,V} = \{f/g \in \mathcal{O}_{v,V} \mid f/g = f_1/g_1 \text{ with } f_1(v) = 0,\ g_1(v) \neq 0\}$$

is the set of rational functions on V that are defined and equal to 0 at v. Since $\mathcal{O}_{v,V}$ is a localization of the Noetherian integral domain $k[V]$ at a prime ideal, $\mathcal{O}_{v,V}$ is also a Noetherian integral domain. Note also that $\mathcal{O}_{v,V}/\mathfrak{m}_{v,V} \cong k[V]/\mathcal{I}(v) \cong k$ by Proposition 46(5).

Recall that the polynomial maps from V to k are also referred to as the *regular* maps of V to k. This is because these are precisely the rational functions on V that are regular everywhere:

Proposition 51. If V is an affine variety over an algebraically closed field k then the rational functions on V that are regular at all points of V are precisely the polynomial functions $k[V]$.

Proof: This follows from Proposition 48, which shows that the intersection (in $k(V)$) of all of the localizations of $k[V]$ at the maximal ideals of $k[V]$ is precisely $k[V]$.

Since the maximal ideals of $k[V]$ are in bijective correspondence with the points of V, the fact that the local ring $\mathcal{O}_{v,V}$ is the same as the localization of $k[V]$ at the maximal ideal corresponding to v shows that $\mathcal{O}_{v,V}$ depends intrinsically on the ring $k[V]$ and is independent of the embedding of V in a particular affine space.

Suppose $\varphi : V \to W$ is a morphism of affine varieties with associated k-algebra homomorphism $\widetilde{\varphi} : k[W] \to k[V]$. If $v \in V$ is mapped to $w \in W$ by φ, then it is straightforward to show that $\widetilde{\varphi}$ induces a homomorphism (also denoted by $\widetilde{\varphi}$) between the corresponding local rings:

$$\widetilde{\varphi} : \mathcal{O}_{w,W} \to \mathcal{O}_{v,V} \quad \text{where} \quad \widetilde{\varphi}(h/k) = \widetilde{\varphi}(h)/\widetilde{\varphi}(k),$$

and that under this homomorphism, $\widetilde{\varphi}^{-1}(\mathfrak{m}_{v,V}) = \mathfrak{m}_{w,W}$ (a homomorphism of local rings having this property is called a *local homomorphism*). Note that $\widetilde{\varphi}$ does not in general extend to a field homomorphism from *all* of $k(W)$ into $k(V)$ since elements of $k[W]$ lying in the kernel of $\widetilde{\varphi}$ do not map to invertible elements in $k(V)$. It is also easy to check that if $\psi \circ \varphi$ is a composition of morphisms then on the local rings $\widetilde{\psi \circ \varphi} = \widetilde{\varphi} \circ \widetilde{\psi}$.

The local ring $\mathcal{O}_{v,V}$ can be used to provide an algebraic definition of the "smoothness" (in the sense of the existence of tangents) of V at v, as we now indicate. Suppose first that $V = \mathcal{Z}(f)$ is the hypersurface variety in $\mathbb{A}^n$ defined by the zeros of an irreducible polynomial f in $k[x_1, \dots, x_n]$. For any point $v = (v_1, \dots, v_n)$ on V let $D_v(f)(x_1, \dots, x_n)$ be the linear polynomial:

$$D_v(f)(x_1, \dots, x_n) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(v)\, x_i,$$

where the partial derivative of f with respect to x_i is given by the usual formal rule for the derivative of a polynomial in x_i (with all other variables considered constant). The polynomial $D_v(f)(x_1 - v_1, \dots, x_n - v_n)$ is the first order Taylor polynomial of the function f at v, so gives the best linear approximation to $f(x_1, \dots, x_n) \in k[x_1, \dots, x_n]$ at v. It follows that if $\mathbf{T}$ is the linear variety $\mathcal{Z}(D_v(f)(x_1, \dots, x_n))$ consisting of those points where $D_v(f)$ is zero, then the translate $v + \mathbf{T}$ is "tangent" to the hypersurface $\mathcal{Z}(f)$ at v.

Example

Suppose $f = x^2 - y \in k[x, y]$, so that $V = \mathcal{Z}(f)$ is just the parabola $y = x^2$. We have $\partial f/\partial x = 2x$ and $\partial f/\partial y = -1$, which at $v = (3, 9)$ are equal to 6 and -1, respectively. Then

$$D_{(3,9)}(f)(x, y) = 6x - y,$$

and the corresponding linear variety $\mathbf{T}$ is the line $y = 6x$ through the origin. The translate $(3, 9) + \mathbf{T}$ is the usual tangent line to the parabola at $(3, 9)$. The Taylor expansion of $x^2 - y$ at $(3, 9)$ is $x^2 - y = [\,6(x - 3) - (y - 9)\,] + (x - 3)^2$. The first order terms are $D_{(3,9)}(f)(x - 3, y - 9)$ and give the best linear approximation to $x^2 - y$ near (3,9).

It is straightforward to extend these notions to any affine variety V in $\mathbb{A}^n$.

Definition. Define the *tangent space to V at v* to be the linear variety

$$\mathbb{T}_{v,V} = \mathcal{Z}(\{D_v(f)(x_1, \dots, x_n) \mid f \in \mathcal{I}(V)\}).$$

The formal partial derivatives are k-linear and obey the usual product rule for derivatives, so the tangent space may be computed from the generators for $\mathcal{I}(V)$:

$$\text{if} \quad \mathcal{I}(V) = (f_1, f_2, \dots, f_m) \qquad \text{then} \quad \mathbb{T}_{v,V} = \bigcap_{i=1}^{m} \mathcal{Z}(D_v(f_i)).$$

Note that $\mathbb{T}_{v,V}$ is an intersection of vector spaces, so is a vector subspace of k^n.

This definition of the tangent space $\mathbb{T}_{v,V}$, while making apparent the connection with tangents to the variety V, seems to depend on the embedding of V in $\mathbb{A}^n$. In fact the tangent space can be defined entirely in terms of the local ring $\mathcal{O}_{v,V}$, as the next proposition proves.

Proposition 52. Let V be an affine variety over the algebraically closed field k and let v be a point on V with local ring $\mathcal{O}_{v,V}$ and corresponding maximal ideal $\mathfrak{m}_{v,V}$. Then there is a k-vector space isomorphism

$$(\mathbb{T}_{v,V})^* \cong \mathfrak{m}_{v,V}/\mathfrak{m}_{v,V}^2$$

where $(\mathbb{T}_{v,V})^*$ denotes the vector space dual (cf. Section 11.3) of the tangent space $\mathbb{T}_{v,V}$ to V at v.

Proof: Let $(k^n)^*$ denote the n-dimensional vector space dual to k^n. Since each $D_v(f)$ is a linear function, D_v is a linear transformation from $k[x_1, \dots, x_n]$ to $(k^n)^*$.

Let M_v be the maximal ideal in $k[x_1, \dots, x_n]$ generated by the set $x_i - v_i$ for $1 \le i \le n$. The image $M_v/\mathcal{I}(V)$ of M_v in $k[V]$ is the ideal $\mathcal{I}(v)$ of functions on V that are zero at v and $\mathcal{I}(v)^2 = M_v^2 + \mathcal{I}(V)$. Then $\mathcal{O}_{v,V}$ is the localization of $k[V]$ at $\mathcal{I}(v)$; and identifying $\mathcal{I}(v)$ with its image in $\mathcal{O}_{v,V}$ we have $\mathfrak{m}_{v,V} = \mathcal{I}(v)\mathcal{O}_{v,V}$ (Proposition 46(2)). By definition of D_v we have $D_v(x_i - v_i) = x_i$, and since these linear functions form a basis of $(k^n)^*$, it follows that D_v maps M_v surjectively onto $(k^n)^*$. The kernel of D_v consists of the elements of $k[x_1, \dots, x_n]$ whose Taylor expansion at v starts in degree at least 2 and these are just the elements in M_v^2. Hence D_v defines an isomorphism

$$D_v : M_v/M_v^2 \xrightarrow{\sim} (k^n)^*.$$

The tangent space $\mathbb{T}_{v,V}$ is a vector subspace of k^n, so every linear function on k^n restricts to a linear function on $\mathbb{T}_{v,V}$. Composing D_v with this restriction map gives a linear transformation

$$D : M_v \xrightarrow{D_v} (k^n)^* \xrightarrow{\text{res}} (\mathbb{T}_{v,V})^*$$

which is surjective since the individual maps are each surjective. We have already seen that $\mathcal{I}(v)^2 = M_v^2 + \mathcal{I}(V)$, so $\mathcal{I}(v)/\mathcal{I}(v)^2 \cong M_v/(M_v^2 + \mathcal{I}(V))$. It follows by Proposition 46(5) that $\mathfrak{m}_{v,V}/\mathfrak{m}_{v,V}^2 \cong \mathcal{I}(v)/\mathcal{I}(v)^2$. To prove the proposition it is therefore sufficient to show that $\ker D = M_v^2 + \mathcal{I}(V)$, since then

$$\mathfrak{m}_{v,V}/\mathfrak{m}_{v,V}^2 \cong M_v/(M_v^2 + \mathcal{I}(V)) = M_v/\ker D \cong (\mathbb{T}_{v,V})^*.$$

The polynomial f is in $\ker D$ if and only if $D_v(f)$ is zero on $\mathbb{T}_{v,V}$, i.e., if and only if the linear term of the Taylor polynomial of f expanded about v lies in $\mathcal{I}(\mathbb{T}_{v,V})$. Since the linear terms of the functions in $\mathcal{I}(V)$ generate the ideal $\mathcal{I}(\mathbb{T}_{v,V})$, it follows that f is in $\ker D$ if and only if $f - g$ has zero linear term for some g in $\mathcal{I}(V)$. But this is equivalent to $f \in \mathcal{I}(V) + M_v^2$, so $\ker D = \mathcal{I}(V) + M_v^2$, completing the proof of the proposition.

Recall that the *dimension* of a variety V is by definition the transcendence degree of the field $k(V)$ over k. Since each local ring $\mathcal{O}_{v,V}$ has $k(V)$ as its field of fractions, the dimension of V is determined by the transcendence degree over k of the field of fractions of any of its local rings.

Definition. We say V is *nonsingular* at the point $v \in V$ (or v is a *nonsingular point* of V) if the dimension of the k-vector space $\mathbb{T}_{v,V}$ is $\dim V$. Equivalently (by Proposition 52), v is a nonsingular point of V if $\dim_k(\mathfrak{m}_{v,V}/\mathfrak{m}_{v,V}^2) = \dim V$. Otherwise the point v is called a *singular point.* The variety V is *nonsingular* or *smooth* if it is nonsingular at every point.

The geometric picture is that at a nonsingular point v there are as many independent tangents as one would expect: a tangent line on a curve, a tangent plane on a surface, etc.

Whether a variety V is nonsingular at a point v can be determined from properties of the local ring $\mathcal{O}_{v,V}$, namely whether $\dim_k(\mathfrak{m}_{v,V}/\mathfrak{m}_{v,V}^2) = \dim \mathcal{O}_{v,V}$. A local ring having this property is said to be a *regular local ring.* In particular, the notion of singularity does not depend on the embedding of V in a specific affine space. This algebraic interpretation can be used to *define* smoothness for abstract algebraic varieties, where the geometric intuition of tangent planes to surfaces (for example) is not as obvious.

If $f_1, \dots, f_m$ are generators for $\mathcal{I}(V)$ defining V in $\mathbb{A}^n$, then the dimension of V can be determined from a Gröbner basis for $\mathcal{I}(V)$ (cf. Exercise 29). Determining the dimension of the tangent space $\mathbb{T}_{v,V}$ as a vector space over k is a linear algebra problem: this vector space is the set of solutions of the m linear equations $D_v(f_i)(x_1, \dots, x_n) = 0$. If r is the rank of the $m \times n$ matrix of coefficients $\partial f_i/\partial x_j(v)$ of this system of equations, then $\mathbb{T}_{v,V}$ is a vector space of dimension $n-r$. Using this it is not too difficult to establish the following:

1. We have $\dim V \leq \dim_k(\mathbb{T}_{v,V}) \leq n$ for every point v in $V \subseteq \mathbb{A}^n$.

2. The set of singular points of V is a proper Zariski closed subset of V. The set of nonsingular points of V is a nonempty open subset of V; in particular the nonsingular points of V are dense in V (so "most" points of V are nonsingular).

We also state without proof the following result which further relates the local geometry of V to the algebraic properties of the local rings of V:

3. If v is a nonsingular point, then the local ring $\mathcal{O}_{v,V}$ is a Unique Factorization Domain; in particular, $\mathcal{O}_{v,V}$ is integrally closed (cf. Example 3 following Corollary 25).

The variety V is said to be *factorial* if $\mathcal{O}_{v,V}$ is a U.F.D. for every point $v \in V$, and is said to be a *normal* variety if $\mathcal{O}_{v,V}$ is integrally closed for every $v \in V$ (which by Proposition 49 is equivalent to $k[V]$ being integrally closed). By (3) above we have

$$\textit{smooth varieties} \quad \subseteq \quad \textit{factorial varieties} \quad \subseteq \quad \textit{normal varieties.}$$

In general each of the above containments is proper. In the case when V has dimension 1, i.e., V is an *affine curve*, however, these three properties are in fact equivalent: we shall prove later that an irreducible affine curve is smooth if and only if it is normal or factorial (cf. Corollary 13 in Section 16.2). It follows that over an algebraically closed field k,

an irreducible affine curve C is smooth if and only if $k[C]$ is integrally closed.

For any irreducible affine curve C the integral closure, S, of $k[V]$ in $k(V)$ is also the coordinate ring of an irreducible affine curve $\widetilde{C}$. Then S is integral over $k[V]$ and, by Theorem 30 and Corollary 27 it follows that there is a morphism from the smooth curve $\widetilde{C}$ onto C that has finite fibers. The curve $\widetilde{C}$ is called the *normalization* or the *nonsingular model* of C, and one can show that it is unique up to isomorphism. Note how the existence of a smooth curve mapping finitely to C (a problem in "geometry") is solved by the existence of integral closures in ring extensions (a problem in "algebra").

We shall give another characterization of smoothness for irreducible affine curves at the end of Section 16.2.

EXERCISES

As usual R is a commutative ring with 1 and D is a multiplicatively closed set in R.

1. Suppose M is a finitely generated R-module. Prove that $D^{-1}M = 0$ if and only if $dM = 0$ for some $d \in D$.
2. Let I be an ideal in R, let D be a multiplicatively closed subset of R with ring of fractions $D^{-1}R$, and let ${}^c({}^eI) = R$ be the saturation of I with respect to D.
   (a) Prove that ${}^c({}^eI) = R$ if and only if ${}^eI = D^{-1}R$ if and only if $I \cap D \neq \emptyset$.
   (b) Prove that $I = {}^c({}^eI)$ is saturated if and only if for every $d \in D$, if $da \in I$ then $a \in I$.
   (c) Prove that extension and contraction define inverse bijections between the ideals of R saturated with respect to D and the ideals of $D^{-1}R$.
   (d) Let $I = (2x, 3y) \subset \mathbb{Z}[x, y]$. Show the saturation of I with respect to $\mathbb{Z} - \{0\}$ is (x, y).
3. If I is an ideal in the commutative ring R let $\varphi : R[x_1, \dots, x_n] \cong (R/I)[x_1, \dots, x_n]$ be the ring homomorphism with kernel $I[x_1, \dots, x_n]$ given by reducing coefficients modulo I. If $\overline{A}$ is an ideal in $(R/I)[x_1, \dots, x_n]$, let A denote the inverse image of $\overline{A}$ under φ.
   (a) For any $i \geq 1$ show that the inverse image under φ of the subring $(R/I)[x_1, \dots, x_i]$ is $R[x_1, \dots, x_i] + I[x_1, \dots, x_n]$.
   (b) Prove that $\varphi(A \cap R[x_1, \dots, x_i]) = \overline{A} \cap (R/I)[x_1, \dots, x_i]$
4. Let $f = y^5 - z^4$, viewed as a polynomial in y with coefficients in $\mathbb{Q}[z]$.
   (a) Prove that f has no roots in $\mathbb{Q}[z]$.
   (b) Suppose $f = (y^2 + ay + b)(y^3 + cy^2 + dy + e)$. Show that a, b, c, d, e satisfy the system of equations
   $$a + c = 0, \quad ac + b + d = 0, \quad ad + bc + e = 0, \quad ae + bd = 0, \quad be - z^4 = 0.$$
   Deduce that $e^5 = z^{12}$ and conclude that f is irreducible in $\mathbb{Q}[y, z]$. [Use elimination.]

5. Suppose R is a U.F.D. with field of fractions F and $p \in R[x]$ is a monic polynomial.

(a) Show that the ideal $pR[x]$ generated by p in $R[x]$ is prime if and only if the ideal $pF[x]$ generated by p in $F[x]$ is prime. [Use Gauss' Lemma.]

(b) Show that $pR[x]$ is saturated, i.e., that $pF[x] \cap R[x] = pR[x]$.

6. Show that $I = (y^3 - xz, xy^2 - z^2)$ is not a prime ideal in $\mathbb{Q}[x, y, z]$ and find explicit elements $a, b \in \mathbb{Q}[x, y, z]$ with $ab \in I$ but $a \notin I$ and $b \notin I$.

7. Show that $P = (y^3 - xz, xy^2 - z^2, x^2 - yz)$ is a prime ideal in $\mathbb{Q}[x, y, z]$.

8. Show that $P = (x^2 - yz, w^2 - x^4z)$ is a prime ideal in $\mathbb{Q}[x, y, z, w]$.

9. Show that $P = (xz^2 - w^3, xw^2 - y^4, y^4z^2 - w^5)$ is a prime ideal in $\mathbb{Q}[x, y, z, w]$.

10. Show that $I = (xy - w^3, y^2 - zw)$ is not a prime ideal in $\mathbb{Q}[x, y, z, w]$ and find a, b with $ab \in I$ but $a, b \notin I$.

11. Let R_P be the localization of R at the prime P. Prove that if Q is a P-primary ideal of R then $Q = {}^c({}^eQ)$ with respect to the extension and contraction of Q to R_P. Show the same result holds if Q is P'-primary for some prime P' contained in P.

12. Let $R = \mathbb{R}[x, y, z]/(xy - z^2)$, let $P = (\bar{x}, \bar{z})$ be the prime ideal generated by the images of x and y in R, and let R_P be the localization of R at P. Prove that $P^2R_P \cap R = (\bar{x})$ and is strictly larger than P^2.

13. Prove that if N and N' are two R-submodules of an R-module M with $N_P = N'_P$ in the localization M_P for every prime ideal P of R (or just for every maximal ideal) then $N = N'$.

14. Suppose $\varphi : M \to N$ is an R-module homomorphism. Prove that φ is injective (respectively, surjective) if and only if the induced R_P-module homomorphism $\varphi : M_P \to N_P$ is injective (respectively, surjective) for every prime ideal P of R (or just for every maximal ideal of R).

15. Let $R = \mathbb{Z}[\sqrt{-5}]$ be the ring of integers in the quadratic field $\mathbb{Q}(\sqrt{-5})$ and let I be the prime ideal $(2, 1 + \sqrt{-5}\,)$ of R generated by 2 and $1 + \sqrt{-5}$ (cf. Exercise 5, Section 8.2). Recall that every nonzero prime ideal P of R contains a prime $p \in \mathbb{Z}$.

(a) If P is a prime ideal of R not containing 2 prove that $I_P = R_P$.

(b) If P is a prime ideal of R containing 2 prove that $P = I$ and that $I_P = (1+\sqrt{-5}\,)R_P$.

(c) Prove that $I_P \cong R_P$ as R_P-modules for every prime ideal P of R but that I and R are not isomorphic R-modules. (This example shows that it is important in Exercise 14 to be *given* the R-module homomorphism φ.) [Observe that $I \cong R$ as R-modules if and only if I is a *principal* ideal.]

16. Prove that localization commutes with tensor products: there is a unique isomorphism of $D^{-1}R$-modules $\varphi : (D^{-1}M) \otimes_{D^{-1}R} (D^{-1}N) \cong D^{-1}(M \otimes_R N)$ with $\varphi((m/d) \otimes (n/d'))$ given by $(m \otimes n)/dd'$ for any R-modules M, N, and multiplicatively closed set D in R.

17. Prove that the R-module A is a flat R-module if and only if A_P is a flat R_P-module for every prime ideal P of R (or just for every maximal ideal of R). [Use Proposition 41, Exercises 14 and 16, and the exactness properties of localization.]

18. In the notation of Example 2 following Corollary 37, prove that $R_f \cong R[x]/(fx - 1)$ if f is not nilpotent in R. [Show that the map $\varphi : R[x] \to R_f$ defined by $\varphi(r) = r/1$ and $\varphi(x) = 1/f$ gives a surjective ring homomorphism and the universal property in Theorem 36 gives an inverse.]

19. Prove that if R is an integrally closed integral domain and D is any multiplicatively closed subset of R containing 1, then $D^{-1}R$ is integrally closed.

20. Suppose that R is a subring of the ring S with $1 \in R$ and that S is integral over R. If D is any multiplicatively closed subset of R, prove that $D^{-1}S$ is integral over $D^{-1}R$.

21. Suppose $\varphi : R \to S$ is a ring homomorphism and D' is a multiplicatively closed subset of S. Let $D = \varphi^{-1}(D')$. Prove that D is a multiplicatively closed subset of R and that the map $\varphi' : D^{-1}R \to D'^{-1}S$ given by $\varphi'(r/d) = \varphi(r)/\varphi(d)$ is a ring homomorphism.

22. Suppose $P \subseteq Q$ are prime ideals in R and let R_Q be the localization of R at Q. Prove that the localization R_P is isomorphic to the localization of R_Q at the prime ideal PR_Q (cf. the preceding exercise).

23. Let $\varphi : A \to B$ be a homomorphism of commutative rings with $\varphi(1_A) = 1_B$, and let P be a prime ideal of A. Let contraction and extension of ideals with respect to φ be denoted by superscripts c and e respectively. Prove that P is the contraction of a prime ideal in B if and only if $P = (P^e)^c$. [Localize B at $\varphi(A - P)$.]

24. *(The Going-down Theorem)* Let S be an integral domain, let R be an integrally closed subring of S containing 1_S, and let k be the field of fractions of R. Suppose that $P_2 \subseteq P_1$ are prime ideals in R and that Q_1 is a prime ideal in S with $Q_1 \cap R = P_1$. Let S_{Q_1} be the localization of S at Q_1.

(a) Show that $P_2 \subseteq P_2S_{Q_1} \cap R$.

(b) Suppose that $a \in P_2S_{Q_1} \cap R$ and write $a = s/d$ with $s \in P_2S$ and $d \in S$, $d \notin Q_1$. If the minimal polynomial of s over k is $x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ with $a_0, \dots, a_{n-1} \in P_2$ (cf. Exercise 12 in Section 3) show that the minimal polynomial of d over k is $x^n + b_{n-1}x^{n-1} + \cdots + b_1x + b_0$ where $b_i = a_i/a^{n-i}$ and conclude that $b_i \in R$. [Use Exercise 10 in Section 3.]

(c) Show that $a \in P_2$ and conclude that $P_2S_{Q_1} \cap R = P_2$. [Show $a \notin P_2$ implies $b_i \in P_2$ for $i = 0, 1, \dots, n-1$, which would imply $d^n \in P_2S \subseteq P_1S \subseteq Q_1$ and so $d \in Q_1$.]

(d) Prove that $P_2S_{Q_1}$ is contained in a prime ideal P of S_{Q_1} with $P \cap R = P_2$. [Use (c) and the previous exercise for $\varphi : R \to S_{Q_1}$.]

(e) Let $Q_2 = P \cap S$. Prove that $Q_2 \subseteq Q_1$ and that $Q_2 \cap R = P_2$.

(f) Use induction together with the previous result to prove the Going-down Theorem: Theorem 26(4).

25. Let k be an algebraically closed field and let $V = \mathcal{Z}(xz - yw) \subset \mathbb{A}^4$. Prove that the set of points v where $f = \bar{x}/\bar{y} \in k(V)$ is regular is precisely the set of points (x, y, z, w) where $y \neq 0$ or $z \neq 0$. [If $f = \bar{a}/\bar{b}$ show that $ay - bx \in (xz - yw)$ as polynomials in $k[x, y, z, w]$ and conclude that $b \in (y, z)$.] Prove that there is no function $a/b \in k(V)$ with $b(v) \neq 0$ for every v where f is regular.

26. *(Differentials of Morphisms)* Let $\varphi : V \to W$ be a morphism of affine varieties over the algebraically closed field k and suppose $\varphi(v) = w$.

(a) Show that φ induces a linear map from the k-vector space M_w/M_w^2 to the k-vector space M_v/M_v^2, and use this to show that φ induces a linear map $d\varphi$ (called the *differential* of φ) from the k-vector space $\mathbb{T}_{v,V}$ to the k-vector space $\mathbb{T}_{w,W}$.

(b) Prove that if $V \subseteq \mathbb{A}^n$, $W \subseteq \mathbb{A}^m$ and $\varphi = (F_1(x_1, \dots, x_n), \dots, F_m(x_1, \dots, x_n))$ then $d\varphi : \mathbb{T}_{v,V} \to \mathbb{T}_{w,W}$ is given explicitly by

$$(d\varphi)(a_1, \dots, a_n) = (D_v(F_1)(a_1, \dots, a_n), \dots, D_v(F_m)(a_1, \dots, a_n)).$$

[If $g = g(y_1, \dots, y_m)$ show that the chain rule implies

$$\frac{\partial(g \circ \varphi)}{\partial x_i}(v) = \sum_{j=1}^{m} \frac{\partial g}{\partial y_j}(w)\, \frac{\partial F_j}{\partial x_i}(v),$$

so that $D_v(g \circ \varphi)(a_1, \dots, a_n) = D_w(g)(b_1, \dots, b_m)$ where $b_j = D_v(F_j)(a_1, \dots, a_n)$. Then use the fact that $g \circ \varphi \in \mathcal{I}(V)$ if $g \in \mathcal{I}(W)$.]

(c) If $\psi : U \to V$ is another morphism with $\psi(u) = v$, prove that the associated $d(\varphi \circ \psi) : \mathbb{T}_{u,U} \to \mathbb{T}_{w,W}$ is the same as $d\varphi \circ d\psi$.

(d) Prove that if φ is an isomorphism then $d\varphi$ is a vector space isomorphism from $\mathbb{T}_{v,V}$ to $\mathbb{T}_{w,W}$ for every $\varphi(v) = w$.

27. Let $V = \mathbb{A}^1$ and $W = \mathcal{Z}(xz - y^2, yz - x^3, z^2 - x^2y) \subset \mathbb{A}^3$. Let $\varphi : V \to W$ be the surjective morphism $\varphi(t) = (t^3, t^4, t^5)$ (cf. Exercise 26 in Section 1). For each $t \in \mathbb{A}^1$ describe the differential $d\varphi : \mathbb{T}_{t,\mathbb{A}^1} \to \mathbb{T}_{(t^3,t^4,t^5),W}$ in the previous exercise explicitly; in particular prove that $d\varphi$ is an isomorphism of vector spaces for all $t \neq 0$ and is the zero map for $t = 0$. Use this to prove that V and W are not isomorphic.

28. If k is a field, the quotient $k[x]/(x^2)$ is called the *ring of dual numbers* over k. If V is an affine algebraic set over k, show that a k-algebra homomorphism from $k[V]$ to $k[x]/(x^2)$ is equivalent to specifying a point $v \in V$ with $\mathcal{O}_{v,V}/\mathfrak{m}_{v,V} = k$ (called a *k-rational point* of V) together with an element in the tangent space $\mathbb{T}_{v,V}$ of V at v.

29. (*Computing the dimension of a variety*) Let P be a prime ideal in $k[x_1, \dots, x_n]$, set $P_0 = 0$ and let $P_i = P \cap k[x_1, \dots, x_i]$. Define the varieties $V_i = \mathcal{Z}(P_i) \subseteq \mathbb{A}^i$ with V_0 the zero dimensional variety consisting of a single point and coordinate ring k.

(a) Show that $\dim V_{i-1} \leq \dim V_i \leq \dim V_{i-1} + 1$. [First exhibit an injection from $k[V_{i-1}]$ into $k[V_i]$; then show that $k[V_i]$ is a k-algebra generated by $k[V_{i-1}]$ and one additional generator.]

(b) If the ideal generated by P_{i-1} in $k[x_1, \dots, x_i]$ equals P_i, show that $V_i \cong V_{i-1} \times \mathbb{A}^1$ and deduce that $\dim V_i = \dim V_{i-1} + 1$.

(c) If the ideal generated by P_{i-1} in $k[x_1, \dots, x_i]$ is properly contained in P_i, show that $\dim V_i = \dim V_{i-1}$.

(d) Show that $\dim V$ equals the number of $i \in \{1, 2, \dots, n\}$ such that the ideal generated by P_{i-1} in $k[x_1, \dots, x_i]$ equals the ideal P_i. Deduce that if G is the reduced Gröbner basis for P with respect to the lexicographic monomial ordering $x_n > \cdots > x_1$ and $G_i = G \cap k[x_1, \dots, x_i]$ where $G_0 = \emptyset$, and N is the number of i with $G_i \neq G_{i-1}$ for $1 \leq i \leq n$, then $\dim V = n - N$.

The following eleven exercises introduce the notion of the *support* of an R-module M and its relation to the associated primes of M. Cf. also Exercises 29 to 35 in Section 1 and Exercises 25 to 30 in Section 5.

Definition. If M is an R-module, then the set of prime ideals P of R for which the localization M_P is nonzero is called the *support* of M, denoted $\text{Supp}(M)$.

30. Prove that $M = 0$ if and only if $\text{Supp}(M) = \emptyset$. [Use Proposition 47.]

31. If $0 \to L \to M \to N \to 0$ is an exact sequence of R-modules, prove that the localization M_P is nonzero if and only if one of the localizations N_P and L_P is nonzero and deduce that $\text{Supp}(M) = \text{Supp}(L) \cup \text{Supp}(N)$. In particular, if $M = M_1 \oplus \cdots \oplus M_n$ prove that $\text{Supp}(M) = \text{Supp}(M_1) \cup \cdots \cup \text{Supp}(M_n)$.

32. Suppose $P \subseteq Q$ are prime ideals in R and that M is an R-module. Prove that the localization of the R-module M_Q at P is the localization M_P, i.e., $(M_Q)_P = M_P$. [Argue directly, or use Proposition 41 and the associativity of the tensor product.]

33. Suppose $P \subseteq Q$ are prime ideals in R and that M is an R-module. Prove that if $P \in \text{Supp}(M)$ then $Q \in \text{Supp}(M)$. [Use the previous exercise.]

34. **(a)** Suppose $M = Rm$ is a cyclic R-module. Prove that $M_P = 0$ if and only if there is an element $r \in R$, $r \notin P$ with $rm = 0$. Deduce that $P \in \mathrm{Supp}(M)$ if and only if P contains the annihilator of m in R (cf. Exercise 10 in Section 10.1).

(b) If $M = Rm_1 + \cdots + Rm_n$ is a finitely generated R-module prove that $P \in \mathrm{Supp}(M)$ if and only if P is contained in $\mathrm{Supp}(Rm_i)$ for some $i = 1, \dots, n$. [Use Proposition 42.] Deduce that $P \in \mathrm{Supp}(M)$ if and only if P contains the annihilator $\mathrm{Ann}(M)$ of M in R. [Note $\mathrm{Ann}(M) = \cap_{i=1}^{n} \mathrm{Ann}(Rm_i)$, then use (a) and Exercise 11 of Section 7.4.]

35. Suppose P is a prime ideal of R with $P \cap D = \emptyset$. Prove that if $P \in \mathrm{Ass}_R(M)$ then $D^{-1}P \in \mathrm{Ass}_{D^{-1}R}(D^{-1}M)$. [Use Proposition 38(3) and Proposition 42.]

36. Suppose $D^{-1}P \in \mathrm{Ass}_{D^{-1}R}(D^{-1}M)$ where $P = (a_1, \dots, a_n)$ is a finitely generated prime ideal in R with $P \cap D = \emptyset$.

(a) Suppose $m/d \in D^{-1}M$ has annihilator $D^{-1}P$ in $D^{-1}R$. Show that $d_i a_i m = 0 \in R$ for some $d_1, \dots, d_n \in D$.

(b) Let $d' = d_1 d_2 \dots d_n$. Show that $P = \mathrm{Ann}(d'm)$ and conclude that $P \in \mathrm{Ass}_R(M)$. [The inclusion $P \subseteq \mathrm{Ann}(d'm)$ is immediate. For the reverse inclusion, show that $b \in \mathrm{Ann}(d'm)$ implies that $b/1$ annihilates m/d in $D^{-1}M$, hence $b/1 \in D^{-1}P$, and conclude $b \in P$.]

37. Suppose M is a module over the Noetherian ring R. Use the previous two exercises to show that under the bijection of Proposition 38(3) the prime ideals P of $\mathrm{Ass}_R(M)$ with $P \cap D = \emptyset$ correspond bijectively with the prime ideals of $\mathrm{Ass}_{D^{-1}R}(D^{-1}M)$.

38. Suppose M is a module over the Noetherian ring R and D is a multiplicatively closed subset of R. Let $\mathcal{S}$ be the subset of prime ideals P in $\mathrm{Ass}_R(M)$ with $P \cap D \neq \emptyset$. This exercise proves that the kernel N of the localization map $M \to D^{-1}M$ is the unique submodule N of M with $\mathrm{Ass}_R(N) = \mathcal{S}$ and $\mathrm{Ass}_R(M/N) = \mathrm{Ass}_R(M) - \mathcal{S}$.

(a) If N' is a submodule of M with $\mathrm{Ass}_R(N') = \mathcal{S}$ and $\mathrm{Ass}_R(M/N') = \mathrm{Ass}_R(M) - \mathcal{S}$ as in Exercise 35 in Section 1, prove that the diagram

$$\begin{tikzcd}
M \arrow[r, "\pi"] \arrow[d, "\varphi"'] & M/N' \arrow[d, "\varphi'"] \\
D^{-1}M \arrow[r, "\pi'"] & D^{-1}(M/N')
\end{tikzcd}$$

is commutative, where π and π' are the natural projections (cf. Proposition 42(6)) and φ, φ' are the localization homomorphisms.

(b) Show that $\mathrm{Ass}_{D^{-1}R}(D^{-1}N') = \emptyset$ and conclude that $D^{-1}N' = 0$ and that π' is injective. [Use the previous exercise, the definition of $\mathcal{S}$, and Exercise 34 in Section 1.]

(c) If x is the kernel K of φ' show that $\mathrm{Ann}(x) \cap D \neq \emptyset$ and that $\mathrm{Ass}_R(K) \subseteq \mathcal{S}$. Show that $\mathrm{Ass}_R(K) \subseteq \mathrm{Ass}_R(M/N')$ implies that $\mathrm{Ass}_R(K) = \emptyset$, and deduce that $K = 0$.

(d) Prove φ and π have the same kernel, i.e., $N = N'$, and this submodule of M is unique.

The next two exercises establish a fundamental relation between the sets $\mathrm{Ass}_R(M)$ and $\mathrm{Supp}(M)$ of prime ideals related to the R-module M.

39. Prove that $\mathrm{Ass}_R(M) \subseteq \mathrm{Supp}(M)$. [If $Rm \cong R/P$ use Proposition 42(4) and Proposition 46(1) to show that $0 \neq (Rm)_P \subseteq M_P$.]

40. Suppose that R is Noetherian and M is an R-module.

(a) If $P \in \mathrm{Supp}(M)$ prove that P contains a prime ideal Q with $Q \in \mathrm{Ass}_R(M)$.

(b) If P is a minimal prime in $\mathrm{Supp}(M)$, show that $P \in \mathrm{Ass}_R(M)$. [Use Exercise 33 in Section 1 to show that $\mathrm{Ass}_{R_P}(M_P) \neq \emptyset$ and then use Exercise 37.]

(c) Conclude that $\mathrm{Ass}_R(M) \subseteq \mathrm{Supp}(M)$ and that these two sets have the same minimal elements.

15.5 THE PRIME SPECTRUM OF A RING

Throughout this section the term "ring" will mean commutative ring with 1 and all ring homomorphisms $\varphi : R \to S$ will be assumed to map 1_R to 1_S.

We have seen that most of the geometric properties of affine algebraic sets V over k can be translated into algebraic properties of the associated coordinate rings $k[V]$ of k-valued functions on V. For example, the morphisms from V to W correspond to k-algebra ring homomorphisms from $k[W]$ to $k[V]$. When the field k is an algebraically closed field this translation is particularly precise: Hilbert's Nullstellensatz establishes a bijection between the points v of V and the maximal ideals $M = \mathcal{I}(v)$ of $k[V]$, and if $\varphi : V \to W$ is a morphism then $\varphi(v) \in W$ corresponds to the maximal ideal $\widetilde{\varphi}^{-1}(M)$ in $k[W]$. In this development we have generally started with geometric properties of the affine algebraic sets and then seen that many of the algebraic properties common to the associated coordinate rings can be defined for arbitrary commutative rings. Suppose now we try to reverse this, namely start with a general commutative ring as the algebraic object and attempt to define a corresponding "geometric" object by analogy with $k[V]$ and V.

Given a commutative ring R, perhaps the most natural analogy with $k[V]$ and V would suggest defining the collection of maximal ideals M of R as the "points" of the associated geometric object. Under this definition, if $\widetilde{\varphi} : R' \to R$ is a ring homomorphism, then $\widetilde{\varphi}^{-1}(M)$ should correspond to the maximal ideal M. Unfortunately, the inverse image of a maximal ideal by a ring homomorphism in general need not be a maximal ideal. Since the inverse image of a *prime* ideal under a ring homomorphism (that maps 1 to 1) *is* prime, this suggests that a better definition might include the prime ideals of R. This leads to the following:

Definition. Let R be a commutative ring with 1. The *spectrum* or *prime spectrum* of R, denoted Spec R, is the set of all prime ideals of R. The set of all maximal ideals of R, denoted mSpec R, is called the *maximal spectrum* of R.

Examples

(1) If R is a field then Spec R = mSpec $R = \{(0)\}$.

(2) The points in Spec $\mathbb{Z}$ are the prime ideal (0) and the prime ideals (p) where $p > 0$ is a prime, and mSpec $\mathbb{Z}$ consists of all the prime ideals of Spec $\mathbb{Z}$ except (0).

(3) The elements of Spec $\mathbb{Z}[x]$ are the following:

(a) (0)

(b) (p) where p is a prime in $\mathbb{Z}$

(c) (f) where $f \neq 1$ is a polynomial of content 1 (i.e., the g.c.d. of its coefficients is equal to 1) that is irreducible in $\mathbb{Q}[x]$

(d) (p, g) where p is a prime in $\mathbb{Z}$ and g is a monic polynomial that is irreducible mod p.

The elements of mSpec $\mathbb{Z}[x]$ are the primes in (d) above.

In the analogy with $k[V]$ and V when k is algebraically closed, the elements $f \in k[V]$ are functions on V with values in k, obtained by evaluating f at the point v in V. Note that "evaluation at v" defines a homomorphism from $k[V]$ to k with kernel $\mathcal{I}(v)$, and that the value of f at v is the element of k representing f in the quotient

$k[V]/\mathcal{I}(v) \cong k$. Put another way, the value of $f \in k[V]$ at $v \in V$ can be viewed as the element $\bar{f} \in k[V]/\mathcal{I}(v) \cong k$. A similar definition can be made in general:

Definition. If $f \in R$ then the *value* of f at the point $P \in \operatorname{Spec} R$ is the element $f(P) = \bar{f} \in R/P$.

Note that the values of f at different points P in general lie in *different* integral domains. Note also that in general $f \in R$ is not uniquely determined by its values, rather f is determined only up to an element in the nilradical of R (cf. Exercise 3).

There are analogues of the maps $\mathcal{Z}$ and $\mathcal{I}$ and also for the Zariski topology. For any subset A of R define

$$\mathcal{Z}(A) = \{P \in X \mid A \subseteq P\} \subseteq \operatorname{Spec} R,$$

the collection of prime ideals containing A. It is immediate that $\mathcal{Z}(A) = \mathcal{Z}(I)$, where $I = (A)$ is the ideal generated by A so there is no loss simply in considering $\mathcal{Z}(I)$ where I is an ideal of R. Note that, by definition, $P \in \mathcal{Z}(I)$ if and only if $I \subseteq P$, which occurs if and only if $f \in P$ for every $f \in I$. Viewing $f \in R$ as a function on $\operatorname{Spec} R$ as above, this says that $P \in \mathcal{Z}(I)$ if and only if $f(P) = f \bmod P = 0 \in R/P$ for all $f \in I$. In this sense, $\mathcal{Z}(I)$ consists of the points in $\operatorname{Spec} R$ at which all the functions in I have the value 0.

For any subset Y of $\operatorname{Spec} R$ define

$$\mathcal{I}(Y) = \bigcap_{P \in Y} P,$$

the intersection of the prime ideals in Y.

Proposition 53. Let R be a commutative ring with 1. The maps $\mathcal{Z}$ and $\mathcal{I}$ between R and $\operatorname{Spec} R$ defined above satisfy

(1) for any ideal I of R, $\mathcal{Z}(I) = \mathcal{Z}(\operatorname{rad}(I)) = \mathcal{Z}(\mathcal{I}(\mathcal{Z}(I)))$, and $\mathcal{I}(\mathcal{Z}(I)) = \operatorname{rad} I$,

(2) for any ideals I, J of R, $\mathcal{Z}(I \cap J) = \mathcal{Z}(IJ) = \mathcal{Z}(I) \cup \mathcal{Z}(J)$, and

(3) if $\{I_j\}$ is an arbitrary collection of ideals of R, then $\mathcal{Z}(\cup I_j) = \cap \mathcal{Z}(I_j)$.

Proof: If P is a prime ideal containing the ideal I then P contains $\operatorname{rad} I$ (Exercise 8, Section 2), which implies $\mathcal{Z}(I) = \mathcal{Z}(\operatorname{rad}(I))$. Since $\operatorname{rad} I$ is the intersection of all the prime ideals containing I (Proposition 12), the definition of $\mathcal{I}(I)$ gives $\mathcal{Z}(\operatorname{rad}(I)) = \mathcal{Z}(\mathcal{I}(I))$. Similarly,

$$\mathcal{I}(\mathcal{Z}(I)) = \bigcap_{P \in \mathcal{Z}(I)} P = \bigcap_{I \subseteq P} P = \operatorname{rad} I,$$

which completes the proof of (1). It is immediate that $\mathcal{Z}(I \cap J) = \mathcal{Z}(I) \cup \mathcal{Z}(J)$. Suppose the prime ideal P contains IJ. If P does not contain I then there is some element $i \in I$ with $i \notin P$. Since $iJ \subseteq P$, it follows that $J \subseteq P$. This proves $\mathcal{Z}(IJ) = \mathcal{Z}(I) \cup \mathcal{Z}(J)$ and completes the proof of (2). The proof of (3) is immediate.

The first statement in the proposition shows that every set $\mathcal{Z}(I)$ in $\operatorname{Spec} R$ occurs for some *radical* ideal I, and since $\mathcal{I}(\mathcal{Z}(I)) = \operatorname{rad} I$, this radical ideal is unique.

The second two statements in the proposition show that the collection

$$\mathcal{T} = \{\mathcal{Z}(I) \mid I \text{ is an ideal of } R\}$$

satisfies the three axioms for the closed sets of a topology on Spec R as in Section 2.

Definition. The topology on Spec R defined by the closed sets $\mathcal{Z}(I)$ for the ideals I of R is called the *Zariski topology* on Spec R.

By definition, the closure in the Zariski topology of the singleton set $\{P\}$ in Spec R consists of all the prime ideals of R that contain P. In particular, a point P in Spec R is closed in the Zariski topology if and only if the prime ideal P is not contained in any other prime ideals of R, i.e., if and only if P is a maximal ideal (so the Zariski topology on Spec R is not generally Hausdorff). These points are given a name:

Definition. The maximal ideals of R are called the *closed points* in Spec R.

In terms of the terminology above, the points in Spec R that are closed in the Zariski topology are precisely the points in mSpec R.

A closed subset of a topological space is *irreducible* if it is not the union of two proper closed subsets, or, equivalently, if every nonempty open set is dense. Arguments similar to those used to prove Proposition 17 show that the closed subset $Y = \mathcal{Z}(I)$ in Spec R is irreducible if and only if $\mathcal{I}(Y) = \text{rad}\, I$ is prime (cf. Exercise 16).

The following proposition summarizes some of these results:

Proposition 54. The maps $\mathcal{Z}$ and $\mathcal{I}$ define inverse bijections

$$\{\text{Zariski closed subsets of Spec } R\} \underset{\mathcal{Z}}{\overset{\mathcal{I}}{\rightleftarrows}} \{\text{radical ideals of } R\}.$$

Under this correspondence the closed points in Spec R correspond to the maximal ideals in R, and the irreducible subsets of Spec R correspond to the prime ideals in R.

Examples

(1) If $X = \text{Spec}\,\mathbb{Z}$ then X is irreducible and the nonzero primes give closed points in X. The point (0) is not a closed point, in fact the closure of (0) is all of X, i.e., (0) is *dense* in Spec $\mathbb{Z}$. For this reason the element (0) is called a *generic point* in Spec $\mathbb{Z}$.

Since every ideal of $\mathbb{Z}$ is principal, the Zariski closed sets in Spec $\mathbb{Z}$ are $\emptyset$, Spec $\mathbb{Z}$ and any finite set of nonzero prime ideals in $\mathbb{Z}$.

(2) Suppose $X = \text{Spec}\,\mathbb{Z}[x]$ as in Example 3 previously. For each integer prime p the Zariski closure of the element $(p) \in X$ consists of the maximal ideals (p, g) of type (d). Likewise for each $\mathbb{Q}$-irreducible polynomial f of type (c), the Zariski closure of the element (f) is the collection of prime ideals of type (d) where g is some divisor of f in $\mathbb{Z}/p\mathbb{Z}[x]$.

Example: (Affine k-algebras)

Suppose $R = k[V]$ is the coordinate ring of some affine algebraic set $V \subseteq \mathbb{A}^n$ over an algebraically closed field k. Then $R = k[x_1, \dots, x_n]/\mathcal{I}(V)$ where $\mathcal{I}(V)$ is a radical ideal in $k[x_1, \dots, x_n]$. In particular R is a finitely generated k-algebra and since $\mathcal{I}(V)$ is radical, R contains no nonzero nilpotent elements.

Definition. A finitely generated algebra over an algebraically closed field k having no nonzero nilpotent elements is called an *affine k-algebra.*

If R is an affine k-algebra, then by Corollary 5 there is a surjective k-algebra homomorphism $\pi : k[x_1, \dots, x_n] \to R$ whose kernel $I = \ker \pi$ must be a radical ideal since R has no nonzero nilpotent elements. Let $V = \mathcal{Z}(I) \subseteq \mathbb{A}^n$. Then $R \cong k[x_1, \dots, x_n]/I = k[V]$ is the coordinate ring of an affine algebraic set over k. Hence *affine k-algebras are precisely the rings arising as the rings of functions on affine algebraic sets over algebraically closed fields.*

By the Nullstellensatz, the points of mSpec R are in bijective correspondence with V, and the points of Spec R are in bijective correspondence with the subvarieties of V. By Theorem 6, morphisms between two affine algebraic sets correspond bijectively with (k-algebra) homomorphisms of affine k-algebras. In the language of categories these results show that over an algebraically closed field k there is an equivalence of categories

$$\left\{ \begin{array}{c} \text{affine algebraic sets} \\ \text{morphisms of algebraic sets} \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{affine } k\text{-algebras} \\ k\text{-algebra homomorphisms} \end{array} \right\}.$$

The map from left to right sends the affine algebraic set V to its coordinate ring $k[V]$. The map from right to left sends the affine k-algebra R to mSpec R. The pair (mSpec R, R) is sometimes called the *canonical model* of the affine k-algebra R.

Over an algebraically closed field k, a k-algebra homomorphism $\varphi : R \to S$ between two affine k-algebras as in the previous example has the property (by the Nullstellensatz) that the inverse image of a maximal ideal in S is a maximal ideal in R. As previously mentioned, one reason for considering Spec R rather than just mSpec R for more general rings is that inverse images of maximal ideals under ring homomorphisms are not in general maximal ideals. When R is an affine k-algebra corresponding to an affine algebraic set V, the space Spec R contains not only the "geometric points" of V (in the form of the closed points in Spec R), but also the non-closed points corresponding to all of the subvarieties of V (in the form of the non-closed points in Spec R, i.e., the prime ideals P of R that are not maximal).

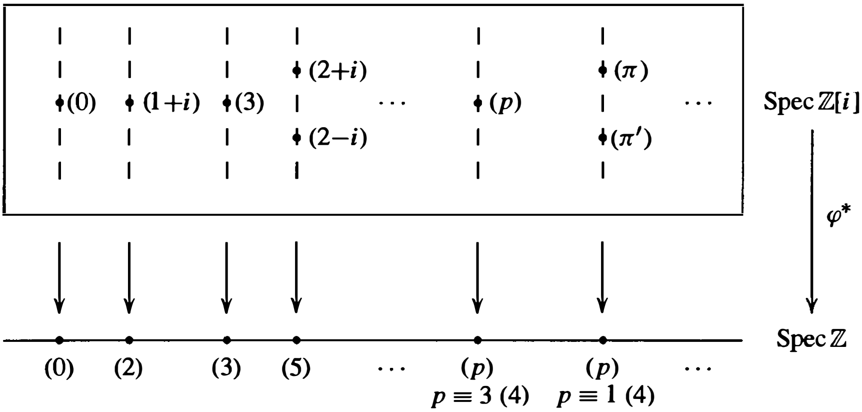
In general, if $\varphi : R \to S$ is a ring homomorphism mapping 1_R to 1_S and P is a prime ideal in S then $\varphi^{-1}(P)$ is a prime ideal in R. This defines a map $\varphi^* :$ Spec $S \to$ Spec R with $\varphi^*(P) = \varphi^{-1}(P)$. If $\mathcal{Z}(I) \subseteq$ Spec R is a Zariski closed subset of Spec R, then it is easy to show that $(\varphi^*)^{-1}(\mathcal{Z}(I))$ is the Zariski closed subset $\mathcal{Z}(\varphi(I)S)$ defined by the ideal generated by $\varphi(I)$ in S. Since the inverse image of a closed subset in Spec R is a closed subset in Spec S, the induced map φ^* is continuous in the Zariski topology. This proves the following proposition.

Proposition 55. Every ring homomorphism $\varphi : R \to S$ mapping 1_R to 1_S induces a map $\varphi^* :$ Spec $S \to$ Spec R that is continuous with respect to the Zariski topologies on Spec R and Spec S.

While the generalization from affine algebraic sets to Spec R for general rings R has made matters slightly more complicated, there are (at least) two very important benefits gained by this more general setting. The first is that Spec R can be considered even for commutative rings R containing nilpotent elements; the second is that Spec R need not be a k-algebra for any field k, and even when it is, the field k need not be algebraically closed. The fact that many of the properties found in the situation of affine k-algebras hold in more general settings then allows the application of "geometric" ideas to these situations (for example, to Spec R when R is finite).

Examples

(1) The natural inclusion $\varphi : \mathbb{Z} \to \mathbb{Z}[i]$ induces a map $\varphi^* : \operatorname{Spec}\mathbb{Z}[i] \to \operatorname{Spec}\mathbb{Z}$. The fiber of φ^* over the nonzero prime P in $\mathbb{Z}$ consists of the prime ideals of $\mathbb{Z}[i]$ containing P. If $P = (p)$ where $p = 2$ or p is a prime congruent to 3 mod 4, then there is only one element in this fiber; if p is a prime congruent to 1 mod 4, then there are two elements in the fiber: the primes (π) and (π') where $p = \pi\pi'$ in $\mathbb{Z}[i]$, cf. Proposition 18 in Section 8.3. This can be represented pictorially in the following figure:



(2) If k is an algebraically closed field then $\operatorname{Spec} k[x]$ consists of (0) and the ideals $(x - a)$ for $a \in k$; the natural inclusion $\varphi : k[x] \to k[x, y]$ induces the Zariski continuous map $\varphi^* : \operatorname{Spec} k[x, y] \to \operatorname{Spec} k[x]$. The elements of $\operatorname{Spec} k[x, y]$ are

(a) (0),

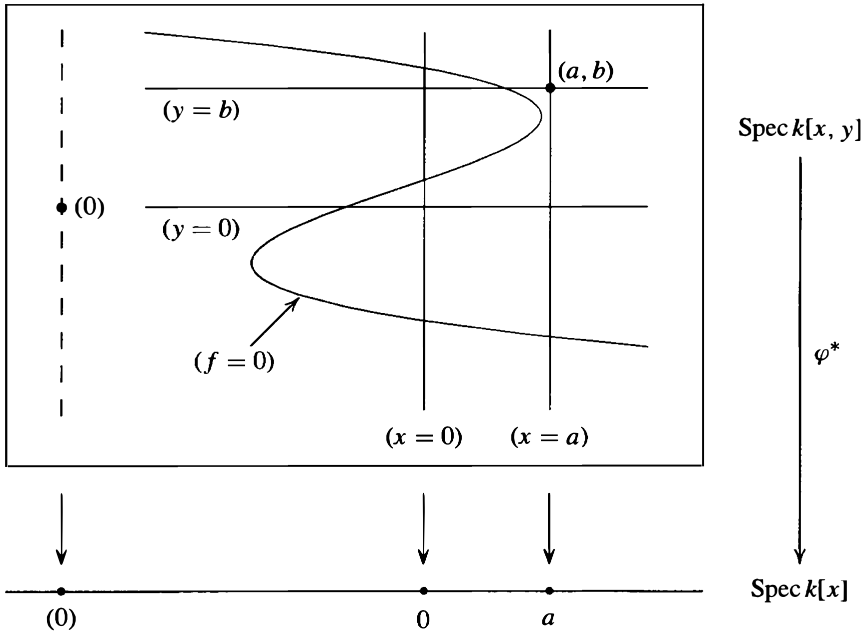
(b) (f) where f is an irreducible polynomial in $k[x, y]$, and

(c) $(x - a, y - b)$ with $a, b \in k$

(cf. Exercise 4). The prime (0) is Zariski dense in $\operatorname{Spec} k[x, y]$; the Zariski closure of the primes in (b) consists of the primes $(x - a, y - b)$ in (c) with $f(a, b) = 0$; the closed points, i.e., the elements of $\operatorname{mSpec} k[x, y]$, are the primes in (c).

By the Nullstellensatz, each prime ideal P in $\operatorname{Spec} k[x, y]$ is uniquely determined by the corresponding zero set $\mathcal{Z}(P)$. The prime $(0) \in k[x, y]$ corresponds to $\mathbb{A}^2$. The prime (f) corresponds to the points where $f(x, y) = 0$, and $P = (f)$ is the intersection of all the maximal ideals containing P. The maximal ideal $(x - a, y - b)$ corresponds to the point $(a, b) \in \mathbb{A}^2$. Fibered over $\operatorname{Spec} k[x]$ by the map φ^* these primes can be pictured geometrically as in the diagram on the following page.

In this diagram, the prime $(x - a)$ in $\operatorname{Spec} k[x]$ is identified with the element $a \in k$. The prime $(x) \in \operatorname{Spec} k[x, y]$ corresponds to the points in $\mathbb{A}^2$ with $x = 0$, i.e.,



with the y-axis in $\mathbb{A}^2$; the prime $(y) \in \operatorname{Spec} k[x, y]$ similarly corresponds to the x-axis. The prime $(f) \in \operatorname{Spec} k[x, y]$ corresponds to the irreducible curve $f(x, y) = 0$ in $\mathbb{A}^2$; the points $(a, b) \in \mathbb{A}^2$ lying on this curve correspond to the maximal ideals $(x - a, y - b) \in \operatorname{Spec} k[x, y]$ containing (f). The closed point $(x - a, y - b) \in \operatorname{Spec} k[x, y]$ corresponds to the "geometric point" $(a, b) \in \mathbb{A}^2$.

Note that $\operatorname{Spec} k[x, y]$ captures all of the geometry of algebraic sets in $\mathbb{A}^2$: every algebraic set in $\mathbb{A}^2$ is the finite union of some subset of the irreducible algebraic sets corresponding to the elements of $\operatorname{Spec} k[x, y]$ pictured above. With the exception of the everywhere dense point (0), the "geometric" picture of $\operatorname{Spec} k[x, y]$ is precisely the usual geometry of the affine plane $\mathbb{A}^2$. When k is not algebraically closed the situation is slightly more complicated, but the picture is similar, cf. Exercise 4.

(3) The situation for $\operatorname{Spec} \mathbb{Z}[x]$, viewed as fibered over $\operatorname{Spec} \mathbb{Z}$ by the natural inclusion $\mathbb{Z} \to \mathbb{Z}[x]$ is very similar to the situation of $\operatorname{Spec} k[x, y]$ in the previous example. The elements of $\operatorname{Spec} \mathbb{Z}[x]$ were discussed in Example 2 following Proposition 54 and can be pictured as in the diagram on the following page.

The element (0) is Zariski dense in $\operatorname{Spec} \mathbb{Z}[x]$. The closure of (p) consists of (p) and all the closed points (p, g) where g is a monic polynomial in $\mathbb{Z}[x]$ that is irreducible mod p. The closure of (f) consists of (f) together with the maximal ideals (p, g) that contain (f), which is the same as saying that the image of f in the quotient $\mathbb{Z}[x]/(p, g)$ is 0, i.e., the irreducible polynomial g is a factor of f mod p. The closed points, $\operatorname{mSpec} \mathbb{Z}[x]$, are the maximal ideals (p, g).

Note that the maximal ideals (p, g) containing (f) are precisely the closed points in $\operatorname{mSpec} \mathbb{Z}[x]$ in the diagram above where the "function" f on $\operatorname{Spec} \mathbb{Z}[x]$ (taking the prime P to $f(P) = f \bmod P \in \mathbb{Z}[x]/P$) is zero. For example, the polynomial $f = x^3 - 4x^2 + x - 9 \in \mathbb{Z}[x]$ fits the diagram above: f is irreducible in $\mathbb{Z}[x]$, and